(Ré)annotation de génomes procaryotes complets - Exploration de groupes de gènes chez les bactéries

Stéphanie Bocs

THESE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

ÉCOLE DOCTORALE LA LOGIQUE DU VIVANT

Spécialité : Génétique

présentée par

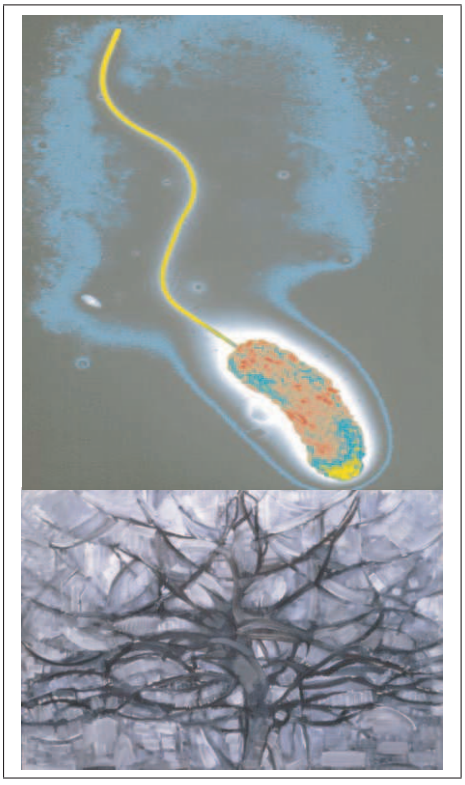
Stéphanie Bocs

Pour obtenir le grade de

Docteur de l'Université Paris 6 – Pierre et Marie Curie

Sujet de la thèse :

(Ré)annotation de génomes procaryotes complets
Exploration de groupes de gènes chez les bactéries



Soutenue le 19 mai 2004 devant le jury composé de :

M.	Pierre	Netter	Président du Jury	Examinateur
Mme.	Claudine	Médigue	Directeur de Thèse	Examinateur
M.	Amos	Bairoch		Rapporteur
M.	Bernard	Prum		Rapporteur
M.	Antoine	Danchin		Examinateur
M.	Jean-Loup	Risler		Examinateur

Remerciements

Je remercie du fond du cœur,

Pierre Netter, directeur de l'école doctorale Logique du Vivant, sans qui cette thèse n'aurait pu voir le jour.

Tous les membres de l'Atelier de Génomique Comparative : Claudine Médigue pour avoir dirigé cette thèse, pour sa disponibilité, ses bons conseils, sa persévérance et son optimisme ! David Vallenet pour ses conseils et son aide sur la modélisation, la gestion, la représentation et l'exploration des annotations. Laurent Labarre pour nos discussions animées sur l'ontologie de l'annotation des génomes procaryotes et pour son travail sur les synténies. Géraldine Pascal pour sa disponibilité, sa joie de vivre qu'elle a su impulser au sein de notre groupe et ses analyses de l'usage des acides aminés dans les séquences protéiques procaryotes. Stéphane Cruveiller pour notre collaboration sur l'apprentissage de différentes classes de gènes. Jérome Le Saux pour son rôle primordial dans le développement, l'utilisation et le blindage du processus de réannotation. Aurélie Lajus, Jean-francois Vincent, Christophe Devine, Angela Jackson et Thibaud Plessis pour leur aide dans le développement d'outils bioinformatiques.

Pierre Tambourin et Hélène Pollard de Genopole pour nous avoir permis d'être hébergé au Centre d'Entreprises et d'Innovation de l'Agglomération d'Evry, le Magellan.

Jean-Loup Risler, Marie-Odile Delorme, Pierre Brezellec, Emmanuelle Ollivier et Claudine Devauchelle du laboratoire Génome et Informatique pour m'avoir donné gout à la bioinformatique et pour nous avoir épaulé dans les moments difficiles.

François Rechenmann, Alain Viari et Stéphane Descorps-Declère du projet Helix (INRIA Rhônes-Alpes), pour le développement des plates-formes de génomiques exploratoires, Imagene et Genostar, *et pour leurs conseils d'expert, notamment sur Prokov et sur les nuées dynamiques.*

Guy Vaysseix, Alexandra Louis, Xavier Benigni, Francis Capy, Didier Gillet, François Laissus, Sylvie Houssay du centre de ressources Infobiogen pour nous avoir accueilli et aidé à administrer nos machines pendant presque deux ans (juillet 2000 - Mars 2002).

Bernard Prum, Francois Képès, Grégory Nuel, Florence Mury, Marie-Hélène Mucchielli du laboratoire Statistiques et Génomes, pour leurs connaissances, leurs conseils avisés, leurs relectures, en particulier, sur les méthodes de prédictions de gènes par contenu.

Jean Weissenbach, Francis Quétier, Claude Scarpelli, Claire Jubin, Yann Esnault, Laurent Sainte-Marthe, Sylvain Bonneval, Claude Verdier-Discala, Valérie Barbe, Véronique de Berardinis, Monique Meugnier, Catherine Sarlande, Corinne Kopec, Susan Cure du Genoscope pour nous avoir intégré au sein de l'UMR Structure et Evolution des Génomes.

Hélène Chiapello et Pierre Nicolas du laboratoire Mathématique, Informatique et Génome, la première, pour son expertise sur l'analyse factorielle de correspondances de l'usage des codons synonymes et le second, pour son expertise sur les modèles de chaînes de Markov Cachées.

Antoine Danchin, Elisabeth Carniel et Eric Duchaud de l'Institut Pasteur pour leurs connaissances en génétique bactérienne, sur les Bacillus, les Yersinia et les Photorhabdus, respectivement.

Eduardo Rocha de l'Atelier de BioInformatique pour ses connaissances en génomique bactérienne, en particulier sur la recherche de répétitions.

Amos Bairoch, Anne-lyse Veuthey, Alexandre Gattiker, Claudia Sapsezian et les annotatrices de l'Institut Suisse de Bioinformatique pour nous avoir intégré au sein du projet HAMAP.

Sandrine Blazy de l'Institut Informatique Electronique pour m'avoir aidé à optimiser l'algorithme de filtrage des CDS.

Patrick Durand d'Hybrigenics pour avoir relu les chapitres ressources et plates-formes de ma thèse.

Patrick Forterre et Laurence Meslet de l'Institut de Génétique et Microbiologie pour leurs connaissances en microbiologie.

Vladimir Pelicic de l'unité de Pathogénie des Infections Systémiques pour ses connaissances sur les Neisseria.

Jacquie Keer et Huw Williams du département de biologie d'Imperial College pour m'avoir donné goût à la recherche.

Gilles Waksman, Alain Bernot, Yveline Kerdaffrec, Nathalie Duong, Florence Hervy pour m'avoir intégré comme monitrice au sein de leur équipe pédagogique du département de biologie de l'université Evry Val d'Essonne.

Alain Sarfati et Sandrine Don de l'Antenne d'Orsay – CIES de Versailles pour m'avoir formé à l'enseignement.

Patrick Faure de Genopole Entreprises pour avoir suivi mon projet lors de ma participation au concours Anvar.

Sophie Gobillard, Francis Arnould Laurent, Santiago Jacques Alès Bianchetti et Aude Decelle pour leur amitié impérissable.

Un merci particulier à la famille qui m'a régulièrement encouragée et soutenue moralement.

Résumé

Au cours de cette thèse, des outils d'annotation et d'exploration des génomes procaryotes ont été développés, validés sur des annotations de référence de bactéries modèles, et appliqués à d'autres organismes.

La base multi-génome PkGDB permet d'intégrer à la fois des ressources hétérogènes et distribuées, et les résultats de diverses méthodes de prédiction. Deux systèmes complémentaires, adossés à PkGDB, sont utilisés pour exploiter les annotations : (i) l'interface Web cartographique et dynamique, *MaGe*, qui a l'avantage de représenter les résultats de synténies, et (ii) la plate-forme *Genostar*, dont les différents modules reposent sur un système de représentation de connaissances factuelles et méthodologiques par objets. La stratégie experte semi-automatique de prédiction de Séquences CoDantes (CDS) d'un chromosome procaryote est fondée sur le modèle statistique des chaînes de Markov. Elle est constituée de deux phases : (i) apprentissage de l'hétérogénéité de composition des CDS du chromosome avec *AMIMat* et (ii) reconnaissance et filtrage des CDS les plus probables avec *AMIGene*. *AMIMat* permet de construire k matrices de transition à partir de k classes de gènes définies selon l'usage des codons synonymes, et à partir d'un jeu de séquences non-codantes natives. La précision d'*AMIGene* dépend de la qualité des k matrices et de la valeur d'autres paramètres validés automatiquement par rapport à des annotations de référence. Le processus de réannotation d'un chromosome procaryote complet permet de comparer le jeu de CDS annotées dans les banques, au jeu de CDS prédites par *AMIGene*, et d'attribuer un statut de réannotation à certaines CDS uniques en fonction de plusieurs critères (longueur, probabilité moyenne de codage, recouvrements, similitudes). L'interface *MaGe* permet alors d'explorer les CDS uniques aux banques, ayant le statut 'faux' ou 'suspect' dans PkGDB, et les CDS uniques à *AMIGene*, ayant le statut 'nouveau' ou 'ambigu'.

Ce processus de réannotation a initialement été testé sur 26 génomes procaryotes, ce qui a conduit à l'intégration de nouveaux polypeptides dans la banque Swiss-Prot. Il est utilisé dans de nombreux projets : *Bacillus*DB, *Entero*DB, *Neisseria*DB, *Acineto*DB, etc. De plus, la présence du gène *secE*, prédit chez *Helicobacter pylori* 26695 et *H. pylori* J99, a été confirmée expérimentalement chez 7 isolats d'*H. pylori*, ce qui a permis de conclure à l'expression de la translocase SecE, composant essentiel de la machinerie de sécrétion. Par la suite, quatre classes de gènes ont été mises en évidence lors de la réannotation des génomes de *Mycobactérium tuberculosis H37Rv* (G+C riche) et de *Photorhabdus luminescens* TT01 (G+C pauvre), alors que seulement trois classes avaient été définies chez *Bacillus subtilis 168* (G+C pauvre) et *Escherichia coli K12* (G+C moyen).

Enfin, nous avons exploré les îlots génomiques d'entérobactéries telles qu'*E. coli* K-12 (commensale), *Photorhabdus luminescens* (entomopathogène) et *Yersinia pestis* CO92 (agent de la peste), dans la perspective de développer une stratégie de prédiction d'îlots génomiques (*e.g.* îlots de pathogénie). L'ensemble de ces outils de génomique bactérienne suscite un grand intérêt tant du point du vue fondamental, où l'on cherche à comprendre les mécanismes du transfert horizontal et son rôle dans l'évolution et la biodiversité, que du point de vue appliqué, où la résistance aux antibiotiques et les infections nosocomiales sont des problèmes de santé publique majeurs.

Table des matières

Table des figures

Liste des tableaux

Préambule

Préambule

Naissance de la génétique moléculaire bactérienne

La génétique moléculaire est marquée de quelques dates importantes. Nous en reprenons ici quelques-unes des plus notables, afin de donner au lecteur un aperçu chronologique de la naissance de la génétique moléculaire bactérienne.

année	découverte	personne	profession	nationalité
1685	premières observation au microscope de bactéries	A. Van Leeuwenhoek[1](1632-1723)	marchant de tissus	hollandais
1859	première théorie de l'évolution	C. Darwin[2](1809-1882)	naturaliste	anglais
1865	lois fondamentales de la génétique	G. Mendel[3](1822-1884)	moine	tchèque
1885	microbiologie	L. Pasteur[4](1822-1895)	chimiste, physicien	français
1900	théorie chromosomique de l'hérédité, et liaison génétique	T. H. Morgan[5](1866-1945)	biologiste	américain
1928	nature biochimique du matériel génétique	F. Griffith[6](1877-1941)	médecin	anglais
1944	identification de l'ADN en tant que support de l'information génétique	O. Avery[7](1877-1955)	médecin	américano-canadien
1943	bactérie comme modèle privilégié de la génétique moderne	M. Delbrück[8](1906-1981), S. E. Luria (1912-1991)	physiciens	allemand, italien
1953	structure de l'ADN	F. H. C. Crick[9](né en 1916), J. D. Watson (né en 1928)	physicien, biologiste	anglais, américain
1965	Régulation de l'expression génétique bactérienne (messager et opéron)	F. Jacob[10](né en 1920), A. Lwoff (1902-1994), J. Monod (1910-1976)	médecin et biologiste, biochimiste, biologiste	français, français, français
1977	premier séquencage d'un fragment d'ADN	P. Berg[11](né en 1926), W. Gilbert (1932), F. Sanger (né en 1932)	chimiste, physicien, chimiste	anglais

[1] http://www.ucmp.berkeley.edu/history/leeuwenhoek.html
[2] http://www.talkorigins.org/faqs/origin.html
[3] http://www.genomenewsnetwork.org/timeline/1866_Mendel.shtml
[4] http://www.pasteur.fr/infosci/biblio/bibliogr/pasteur.html
[5] http://www.nobel.se/medicine/laureates/1933/
[6] http://www.genoscope.cns.fr/externe/HistoireBM/\#griffith
[7] http://profiles.nlm.nih.gov/CC/
[8] http://www.nobel.se/medicine/laureates/1969/
[9] http://www.nobel.se/medicine/laureates/1962/
[10] http://www.nobel.se/medicine/laureates/1965/
[11] http://www.nobel.se/chemistry/laureates/1980/

Naissance de la bioinformatique

Tout comme pour la génétique moléculaire, nous relevons ici quelques dates marquantes dans la naissance de la bioinformatique.

année	découverte	personne	profession	nationalité
1854	algèbre de Boole, fondement théorique des opérations binaires	G. Boole[12](1815-1864)	mathématicien	anglais
1858	premier cable télégraphique[13]	-	-	u.s.a
1934	machine Von Neumann, fondement des ordinateurs à arithmétique binaire	J. Von Neumann[14](1903-1957)	mathématicien	américano-hongrois
193_	fondements théoriques de l'informatique	A. Turing[15](1912-1954), K. Gödel[16](1906-1978), A. Church[17](1903-1976)	mathématiciens	anglais, tchèque, américain
1947	premier calculateur électronique	J. Presper Eckert[18](né en 1919)	ingénieur électricien	américain
1949	théorie de la communication et de l'information	C. Shannon[19](1916-2001)	mathématicien	américain
1956	langage *FORTRAN*, premier langage procédural de haut niveau	J. Backus[20](né en 1924)	mathématicien	américain
1962	théorie de l'horloge moléculaire[21], fondement de l'évolution et de la phylogénie moléculaires	L. Pauling[22](1901-1994), E. Zuckerkandl	chimiste	américain
1963	ouverture du Centre de Recherche en Biochimie Macromoléculaire de Montpellier [Haiech, 2002]	E. Zuckerkandl[23](né en 1922)	biologiste	américano-autrichien
1965	premières comparaisons de séquences protéiques	M. Dayhoff[24] (1925-1983)	biochimiste	américaine
1982	protocole TCP-IP[25]	département de la défense	-	usa

Présentation de la discipline, des thèmes de recherche et du plan de thèse

Le premier génome entièrement séquencé est celui du bactériophage phiX174 (5386 pb) en 1977 [Sanger *et al.*, 1977]. Les 17 années suivantes ont vu apparaître plusieurs autres génomes de virus, de mitochondries et de chloroplastes. En 1995, la première séquence complète du génome de la bactérie *Haemophilus influenzae* et les annotations de cette séquence sont publiées par le groupe américain TIGR[25] [Fleischmann *et al.*, 1995]. *H. influenzae* marque donc le début de l'âge de la génomique bactérienne. La séquence de *H. influenzae* (chromosome circulaire de 1 830 kb) inclut environ 1743 gènes putatifs, parmi lesquels 40% n'ont pas de fonction connue. La moitié de ces gènes n'ont pas d'homologues dans les bases de données, alors que l'autre moitié possède des homologues dont on ignore aussi la fonction. Cette observation s'est répétée à chaque publication d'un nouveau génome, même si les chiffres exacts varient suivant l'organisme et les méthodes utilisées. Une actualisation des annotations de *H. influenzae* a permis d'assigner une fonction à 15% de ces *gènes orphelins* qui, parfois, résultaient d'erreurs de séquençage. Le quatrième génome procaryote a avoir été entièrement séquencé est celui de *Methanococcus jannaschii*, une archeae avec un chromosome circulaire de 1 664 kb et deux plasmides de 58 Kb et 16 Kb [Bult *et al.*, 1996]. Cette archaebactérie vit dans des conditions extrêmes (94°C et 200 atmosphères). Elle est anaérobie stricte et méthanogène : elle produit du méthane en réduisant le CO_2 avec le H_2 pour sa production d'énergie. Sur les 1738 gènes prédits, seuls 38% des gènes ont pu se voir attribuer une fonction précise, ce qui illustre de façon saisissante notre ignorance du domaine des archaea. Le premier génome eucaryote complètement séquencé est celui de *Sacharomyces cerevisiae* [Goffeau *et al.*, 1996]. Ce génome est remarquablement compact pour un génome Eucaryote, puisqu'il possède 16 chromosomes dans un ensemble de 12 Mb et environ 72% de régions codantes. *S. cerevisiae* possède environ 6 200 gènes putatifs, parmi lesquels 30 à 35% n'avaient pas d'homologues dans les bases de données. Il a fallu attendre fin 1997 pour pouvoir enfin accéder aux génomes complets des deux principaux modèles bactériens : *E. coli* K-12 pour les *Protéobactéries* (à Gram négatives) et *Bacillus subtilis* pour les *Firmicutes* (à Gram positives). En Avril 2004, on recense 186 génomes complets, dont 18 archaea,

[12] `http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Boole.html`
[13] `http://www.infobiogen.fr/services/deambulum/fr/bioinfo_hist.html`
[14] `http://ei.cs.vt.edu/~history/VonNeumann.html`
[15] `http://www.turing.org.uk/turing/`
[16] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Godel.html`
[17] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Church.html`
[18] `http://americanhistory.si.edu/csr/comphist/eckert.htm`
[19] `http://www-history.mcs.st-and.ac.uk/~history/Mathematicians/Shannon.html`
[20] `http://www.digitalcentury.com/encyclo/update/backus.html`
[21] Le taux de mutation est constant par rapport au temps et donc proportionnel au temps séparant l'émergence de deux espèces.
[22] `http://www.paulingexhibit.org/bio/index.html`
[23] `http://hrst.mit.edu/hrs/evolution/public/profiles/zuckerkandl.html`
[24] `http://www.isoc.org/internet/history/index.shtml`
[25] `http://www.tigr.org/`

142 bactéries parmi lesquelles de très nombreuses bactéries pathogènes, et 26 eucaryotes : arabette, la drosophile, génome humain ... 487 projets de séquençage de génomes Procaryotes et 415 projets de séquençage de génomes Eucaryotes sont en cours.

La biologie est une discipline complexe à l'interface des mathématiques, de la chimie, de la physique et de l'informatique, d'où la naissance de la biostatistique, de la biochimie, de la biophysique et de la bioinformatique. Cette multidisciplinarité nous permet d'utiliser les méthodes de combinatoire[26] statistique et de l'intelligence artificielle au service de la connaissance en génétique moléculaire procaryote. Cette thèse de bioinformatique a débuté en 1999, à l'Atelier de Génomique Comparative, sous la direction de C. Médigue. Ce travail de recherche dont la problématique est *l'exploration des génomes procaryotes complets* s'est déroulé en trois temps : (i) l'identification de gènes, (ii) la caractérisation fonctionnelle de leurs produits et (iii) la reconstruction de groupes de gènes selon une thématique (*e.g.* îlots génomiques acquis par transfert horizontal). Ce manuscrit est découpé en *trois parties*. D'abord, la *première partie* décomposée en *cinq chapitres*, présente les génomes procaryotes et les outils d'annotation :

1. Biologie des génomes procaryotes
2. Ressources disponibles pour l'étude des génomes procaryotes
3. Prédiction de gènes dans les séquences procaryotes et analyse de leur usage des codons
4. Plates-formes bioinformatiques
5. Annotation *in silico* de génomes procaryotes

Ensuite, la *deuxième partie* est consacrée aux développements méthodologiques mis en oeuvre pour (ré)annoter les génomes procaryotes. Enfin, la *troisième partie* présente des résultats de (ré)annotations et de groupes de gènes construits en combinant plusieurs critères. Cet ouvrage commencera donc par présenter une panoplie d'outils d'analyse de séquences procaryotes pour se focaliser finalement sur la thématique des îlots génomiques d'entérobactéries pathogènes.

Sur la page de couverture, l'image du haut représente *Vibrio cholerae* grossi 10000 fois [Waldor & RayChaudl et celle du bas, une vision possible de l'arbre de la vie [Charlebois *et al.*, 2003].

[26] Etude des objets finis discrets en opposition avec les mathématiques continues.

Première partie

Etat de l'art d'annoter *in silico* des génomes procaryotes

Chapitre 1

Biologie des génomes procaryotes

Ce travail de recherche porte sur l'annotation et l'exploration des génomes procaryotes, en particulier de génomes bactériens. Dans ce chapitre, nous allons d'une part définir les connaissances nécessaires à la modélisation des objets génomiques et d'autre part poser les questions fondamentales que nous avons abordées au travers de ce travail. Même si l'ontologie des génomes procaryotes est un vaste sujet, nous nous devons d'en poser les bases.

1.1 Bactéries

En avril 2004, 142 génomes complets bactériens sont disponibles (d'après le site GOLD[1]). La quantité de données expérimentales biologiques (physiologie, biochimie, génétique moléculaire) est relativement réduite comparée à cette grande quantité de séquences. Il faut aussi tenir compte du fait que certaines de ces bactéries sont non cultivables (*Buchnera aphidicola*), poussent très lentement (*Mycoplasma genitalium*, *M. tuberculosis* H37Rv) ou très difficilement en laboratoire (*P. luminescens*, *Methanococcus jannaschii*). Les deux principaux modèles bactériens sont *E. coli* K-12 et *B. subtilis*, génomes à croissance rapide pour lesquels de nombreuses données expérimentales sont disponibles tant du point de vue de leur biochimie que de leur génétique. Nous avons ultérieurement ajouté à ces deux espèces bactériennes *M. tuberculosis* H37Rv, indispensable dans le cadre de ce travail de génomique comparative.

1.1.1 Compartimentation, métabolisme, croissance et physiologie

Nous avons choisi comme référence trois espèces bactériennes, pour des raisons historiques, pratiques, biologiques et phylogénétiques. *E. coli* K-12 et *B. subtilis* sont des micro-organismes non pathogènes, utilisés comme modèles depuis les prémices de la génétique moléculaire bactérienne.

1. *E. coli* K-12, bacille mobile à coloration gram négatif, appartient au phylum des *Protéobactéries* (ordre des *Entérobactéries*). C'est une bactérie commensale de l'homme (son temps de génération est de 20 minutes à 37°C sur milieu complet en phase exponentielle).

[1] `http://www.genomesonline.org/`

2. *B. subtilis*, bacille mobile à coloration gram positif, appartient au phylum des *Firmicutes* (ordre des *Bacilles*). C'est une bactérie saprophyte qui vit essentiellement à la surface des feuilles des plantes, on la trouve donc aussi parfois dans le sol, l'eau et l'air [Danchin & Sekowska, 1993]. Elle présente les mêmes conditions de croissance en laboratoire qu'*E. coli* K-12. Dans certaines conditions (carences alimentaires), cette bactérie est capable de sporuler.
3. *M. tuberculosis* H37Rv appartient au phylum des *Actinobactéries* (ordre des *Actinomycètes*). Ce bacille immobile a une coloration gram positif atypique car elle est entourée d'une paroi très épaisse due à la présence de couches supplémentaires de lipides peu courants (*i.e.* des acides mycoliques). *M. tuberculosis* H37Rv est l'agent de la tuberculose, première cause de mortalité au début du siècle. L'homme représente, pour cette bactérie, à la fois un hôte et un réservoir, puisque cette bactérie possède la capacité d'entrer en phase de latence (le tubercule est différent d'une spore) et d'échapper ainsi au système immunitaire. Contrairement à *E. coli* K-12, c'est une bactérie à croissance lente (temps de génération de 24h).

Au cours de notre travail, nous avons aussi étudié d'autres espèces bactériennes, comme *P. luminescens*. *P. luminescens* est une entérobactérie pathogène d'insectes vivant en symbiose avec un ver (nématode). Cette bactérie, initialement caractérisée par l'équipe de Noël Boemare (INRA-Université), vit dans le tube digestif d'un nématode. Lorsque le ver s'attaque à des larves d'insectes, il crée de petites lésions qui permettent à la bactérie de s'introduire dans l'hémolymphe de l'insecte. Elle sécrète alors tout une gamme de facteurs de virulence entraînant une mort rapide de la proie. La bioconversion du corps de la proie par des enzymes de la bactérie permet à celle-ci de s'y multiplier tandis que le ver se reproduit, et de s'associer de nouveau au nématode avant de quitter le cadavre de l'insecte. *P. luminescens* doit aussi défendre le cadavre de l'insecte des microbes qui entrent en compétition avec elle. Cette bactérie sécrète pour cela des substances capables de détruire d'autres bactéries ou des champignons. *P. luminescens* est ainsi capable d'anéantir, portée par son vecteur, une large variété d'insectes (toxines), de bactéries (antibiotiques) et de champignons (antifongiques). De plus, c'est la seule espèce bactérienne terrestre identifiée qui possède la propriété de bioluminescence. Cette entérobactérie se présente sous la forme d'un bacille motile[2] à gram négatif.

Les bactéries sont des organismes asexués, leur reproduction se fait par division cellulaire. Habituellement, la croissance conduit à la division de la bactérie (appelée cellule mère) en deux bactéries « identiques[3] » (appelées cellules filles). Ainsi chez les bactéries, croissance et reproduction sont étroitement liées, et l'expression « croissance bactérienne » est généralement employée pour désigner les deux processus. La représentation graphique de la croissance bactérienne en milieu liquide

[2]Une bactérie est dite motile si elle capable de faire des rotations sur elle même grâce aux pili (fimbriae). Ces filaments sont présents chez les bactéries gram négatives mais rarement chez les gram positives. Une bactérie est dite mobile si elle est capable de se déplacer grâce à des flagelles (elle peut nager en milieu liquide ou ramper sur un milieu solide).

[3]La reproduction à l'identique est discutable dès qu'on parle de différences entre les deux brins du chromosome (voir p. 42).

riche (logarithme du nombre de bactéries par millilitre de culture au cours du temps) se découpe généralement en trois phases : la phase de latence, la phase exponentielle et la phase stationnaire (plateau où le nombre de nouvelles bactéries équivaut au nombre de bactéries qui meurent). Le métabolisme des bactéries est très actif en phase exponentielle relativement à la phase stationnaire. A l'origine de ce phénomène se trouve une régulation différentielle de l'expression des gènes en fonction de stimulus extérieurs comme les carences nutritives ou la forte densité de population *quorum sensing*. Dans leur environnement naturel, les bactéries passent la majorité de leur temps sans se multiplier (phase stationnaire). Elles présentent une résistance accrue à de nombreux autres stress : la température, le pH, l'osmolarité, les concentrations gazeuses en O_2 et CO_2, la pression, des molécules toxiques (les antibiotiques, les désinfectants comme l'hypochlorite ($HOCL$), le peroxyde d'hydrogène (H_2O_2), les radicaux libres comme l'anion superoxide ($O_2^{\cdot -}$), le radical hydroxyle ($HO^{\cdot}$)).
Un autre exemple de régulation de l'expression génique est celui des espèces mobiles. Elles déclenchent une migration dans une direction préférentielle en fonction d'un gradient de concentration, soit pour se rapprocher d'une zone favorable, soit pour s'éloigner d'une zone toxique (chimiotactisme ou chimiotaxie).

Il existe au moins deux processus fondamentaux à l'origine de la vie : la compartimentation et le métabolisme. Les bactéries possèdent un intérieur (cytoplasme), une membrane cytoplasmique et une paroi qui les protège du milieu extérieur. La paroi des bactéries est constituée d'un élément de base, le peptidoglycane, encore appelé muréine. C'est un polymère complexe appartenant aux glycopeptides, qui peut aussi être vu comme un polyoside, une mucoprotéine ou un antigène. La paroi des bactéries à gram positif est constituée en majorité de peptidoglycane (jusqu'à 90%) alors que celle des bactéries gram négatif n'est constituée que de 5 à 20% de peptidoglycane. Une autre différence fondamentale entre les parois des bactéries à gram positif et négatif est que la paroi des bactéries à gram négatif est aussi constituée d'une membrane externe qui définit le périplasme (l'espace compris entre les deux membranes).

En plus de cette architecture moléculaire constante, il existe d'autres structures inconstantes qui caractérisent les bactéries. On peut citer la présence d'enveloppes supplémentaires comme la capsule (rôle protecteur) ou le glycocalyx (rôle d'adhésion), et d'appendices, comme les flagelles (encore appelés cils ; permettant la locomotion) ou comme les pili (encore appelé fimbriae ; permettant l'adhésion). Au niveau du métabolisme, la réplication (copie de la molécule d'ADN ou du génome), la transcription (les gènes sont transcrits en ARN messagers qui constituent le transcriptome) et la traduction (les ARNm sont traduits en polypeptides qui constituent le protéome) sont des processus communs aux bactéries alors que d'autres métabolismes participent à la spécificité de chaque espèce (acétogènes, méthanogènes, autotrophie, métabolisme des nitrates, des sulfates [Pelmont, 1993]).

Traditionnellement, en termes moléculaires le gène désigne une séquence nucléotidique (ADN ou ARN) nécessaire à la synthèse d'un polypeptide ou bien spécifiant un ARN fonctionnel. Le dogme central de la biologie moléculaire stipulant qu'à un gène correspond un polypeptide n'est donc qu'en partie vrai ($ADN \rightarrow ARN \rightarrow polypeptide$). Outre les gènes d'ARN fonctionnels ($ADN \rightarrow$

ARN), il existe d'autres contre exemples du dogme central, comme les rétroposons (rétrovirus, rétrotransposons $ARN \rightarrow ADN \rightarrow ARN \rightarrow polypeptide$) qui ne suivent pas la voie classique (réplication, transcription, traduction) mais intègrent une étape supplémentaire de transcription inverse. Les polypeptides peuvent se regrouper en complexes protéiques et peuvent avoir un rôle structural (OMPA consolide la membrane externe), de transport (les porines OMPF permettent le passage de petites molécules hydrophiles, en particulier sur le plan médical des antibiotiques) ou de catalyseur de réactions biochimiques (l'enzyme phosphoglucose isomérase catalyse la conversion du glucose-6-P en fructose-6-P, première réaction de la glycolyse).

Partant du catalogue des gènes codants des fonctions enzymatiques, on va aussi chercher à décrire un organisme vivant par l'ensemble des voies métaboliques qui le caractérisent (on parle aujourd'hui du métabolome). Selon le consortium *Gene Ontology* [Ashburner *et al.*, 2000], les fonctions des protéines doivent être définies suivant trois catégories : le processus biologique, la fonction moléculaire et les composants cellulaires. Le processus biologique réfère à l'objectif biologique auquel la protéine contribue. Il existe des processus larges comme la maintenance et la croissance cellulaire, et la transduction de signaux. Il existe des processus plus spécifiques comme la traduction, le métabolisme des pyrimidines, ou encore la biosynthèse de l'AMPc. La fonction moléculaire est définie comme l'activité biochimique (incluant la fixation à des ligands ou à des structures). Il existe des termes fonctionnels larges comme enzyme, transporteur, ou ligand et d'autres plus précis comme adénylate cyclase ou ligand du récepteur Toll. Les composants cellulaires font référence à l'endroit de la cellule où la protéine est active : cytoplasme, membrane, périplasme, etc.

1.1.2 Objets génomiques, génétique et annotations

Un génome est constitué d'un ou plusieurs réplicons (l'unité de réplication, circulaire ou linéaire, possède une origine et un terminus de réplication) qui sont représentés par un ou plusieurs chromosomes et éventuellement un ou plusieurs plasmides. Les organismes procaryotes se distinguent des organismes eucaryotes, plus complexes, par les propriétés suivantes : leurs génomes sont compacts (quelques Mb ; 10% de non-codant contre 90% chez les eucaryotes ; Tab. 1.1 A p. 37, les gènes sont non morcelés et le plus souvent organisés en opérons.

La figure 1.1 (p. 35) représente deux opérons (unité de transcription orientée de $5' \rightarrow 3'$). Un opéron est constitué d'un ou plusieurs cistrons (gènes). Dans le cas d'un opéron codant des polypeptides, le gène est défini comme l'unité de traduction, il est donc inclus dans une phase ouverte de lecture (*open reading frame*, ORF ; Fig. 1.1 A p. 35). Du fait de la nature du code génétique et de la double hélice de bases appariées constituant l'ADN (voir p. 39), il existe trois cadres de lecture des triplets sur chacun des deux brins. Ainsi une ORF est définie comme la séquence comprise entre deux codons de terminaison de la traduction en phase. La CDS (*coding sequence*, unité de codage dont la longueur moyenne est de 950 pb ; Tab. 1.1 A p. 37) est définie par la séquence comprise entre un codon d'initiation et le premier codon de terminaison en phase. Un objet génomique est donc défini par ses positions de début et de fin ainsi que son orientation (directe ou inverse). Finalement, les polypeptides issus de la traduction (unité fonctionnelle) sont

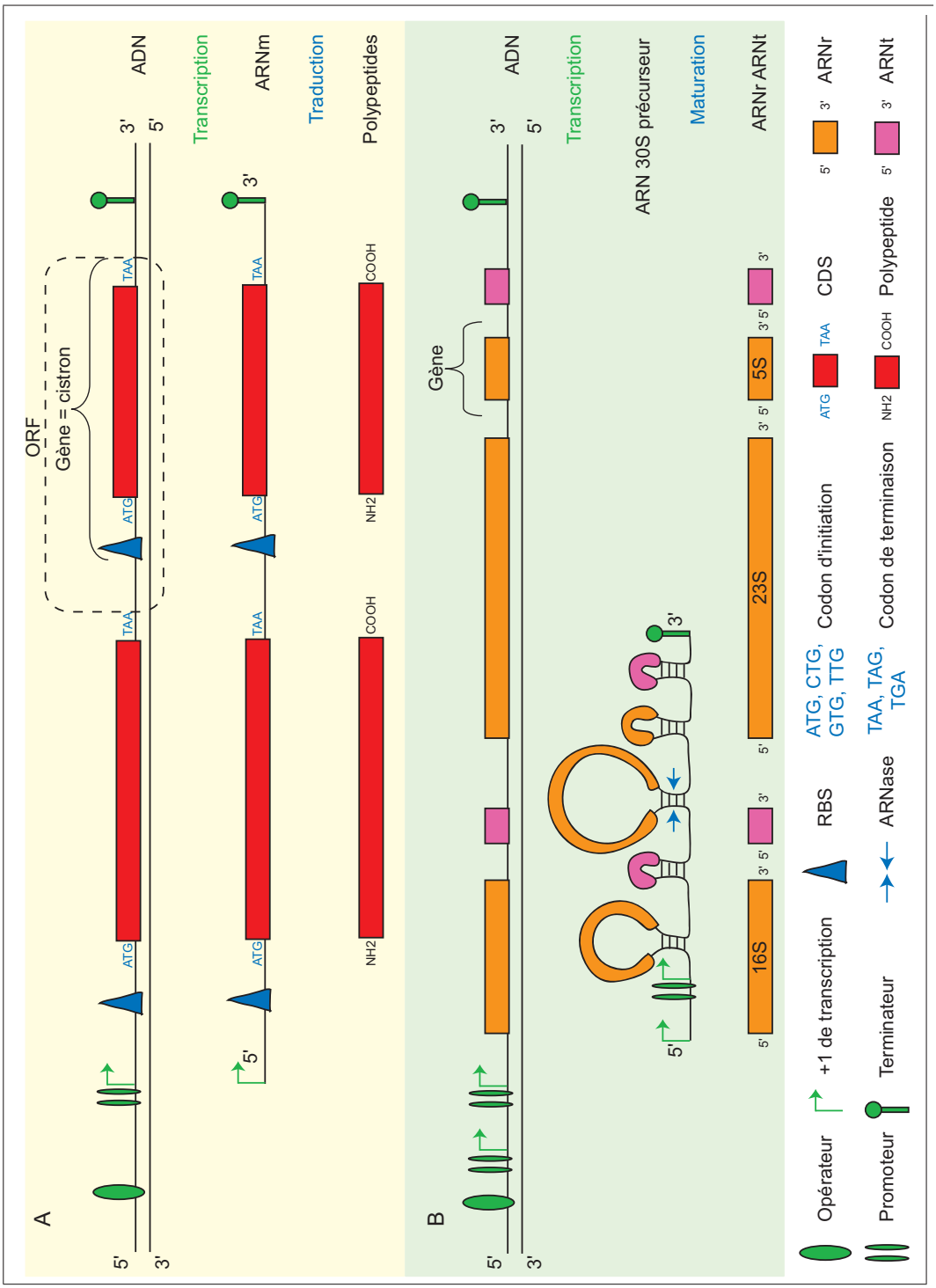


FIG. 1.1 – Objets génomiques

A) Schéma d'un opéron codant des polypeptides concernant l'opéron (unité de transcription), le gène (unité de traduction), la CDS (unité codante), et le polypeptide (unité fonctionnelle). Cet opéron se situe sur le brin direct ($5' \rightarrow 3'$). Dans cet exemple, le brin direct correspond au brin codant et le brin inverse ($3' \rightarrow 5'$), au brin matrice. Autrement dit, l'ARNm est une séquence complémentaire de celle du brin matrice et identique à celle du brin codant. La transcription (resp. la traduction) est effectuée par l'ARN polymérase (resp. le ribosome). L'opérateur et le promoteur (resp. le terminateur rho indépendant) sont des signaux d'initiation de la transcription (resp. de terminaison). Le RBS et le codon d'initiation (resp. le codon de terminaison) sont des signaux d'initiation de la traduction (resp. de terminaison). Le promoteur d'*E. coli* K-12 est constitué de deux boîtes débutant en général aux positions -35 (TTGACA) et -10 (TATAAT) par rapport au +1 de transcription. Le RBS d'*E. coli* K-12 ([AT][CA]AGGA) débute en général à la position -13 par rapport au codon d'initiation. La région qui sépare le RBS du codon d'initiation (espaceur) fait moins de 10 pb. Le terminateur rho indépendant inclut des régions palindromiques qui forment une structure en épingle à cheveux (tige – boucle). La tige inclut une région riche en appariements G-C. La tige boucle (7 à 20 pb) est suivie d'un stretch de U (au moins 3 U).

B) Schéma d'un opéron spécifiant des ARN fonctionnels : l'opéron (unité de transcription), le gène (unité de maturation), l'ARN (unité fonctionnelle). La maturation de l'ARN 30S précurseur en ARN fonctionnels est effectuée par des ARNases. Il existe une différence de quantité transcrite entre les deux promoteurs (P1 et P2 [Keener & Nomura, 1996]).

repliés et assemblés pour former des complexes protéiques grâce aux protéines chaperonnes (DnaK de la famille protéique Hsp70 [Mayhew & F.-U., 1996]).

La régulation de l'expression des gènes peut s'opérer à différents niveaux : transcriptionnel (du promoteur intergénique) et post-transcriptionnel (dégradation de l'ARNm), traductionnel (usage des codons), post-traductionnel (adressage). L'expression des gènes étant régulée de manière très précise, plusieurs signaux de régulation sont aussi recherchés : la région opératrice (sur laquelle se fixe une protéine régulatrice, soit un activateur, soit un répresseur en fonction d'un stimuli extérieur), la région promotrice de la transcription, le site de fixation du ribosome ou RBS (*ribosome binding site*), la région de terminaison de la transcription. L'ensemble de ces objets génomiques ne sont pas toujours simples à caractériser d'un point de vue informatique (voir p. 197), en particulier lorsque qu'il existe des décalages de lecture dans les gènes (ou *frameshift* ; voir p. 119), des recouvrements entre gènes adjacents, entre gènes et signaux ou entre signaux. Par exemple dans le cas des CDS, la détermination du « vrai » codon d'initiation est une tâche difficile car il existe souvent plusieurs codons d'initiations possibles. Ceci est dû à l'ambiguïté des trois triplets [ACT]TG qui peuvent servir à la fois de codon d'initiation (signal de régulation de la traduction) et de codon sens (méthionine, leucine, valine). Une convention simple consiste à choisir systématiquement le codon d'initiation le plus en 5′ mais elle présente au moins deux inconvénients :

1. en voulant résoudre le problème du choix du codon d'initiation, on amplifie les problèmes de chevauchement entre CDS adjacentes qui théoriquement sont au maximum de l'ordre de quelques dizaines de codons
2. la CDS contient alors une partie artefactuelle ce qui peut être gênant par exemple lorsqu'on traduit la séquence pour rechercher des similitudes dans une banque de séquences protéiques.

Généralement, on commence par choisir le codon d'initiation le plus en 5′, puis des méthodes d'analyse aident à le réajuster si nécessaire.

Enfin, il est essentiel d'introduire la notion de régulon qui caractérise un ensemble de gènes sous le contrôle d'un même régulateur. Par exemple, on peut étudier le rôle du régulon RpoS correspondant aux gènes dont l'expression est contrôlée par le facteur sigma RpoS, dans la virulence des Salmonelles et dans leur capacité à résister à certains stress. Un régulon est donc un ensemble d'opérons monocistroniques et/ou polycistroniques, co-régulés.

Les gènes d'un opéron spécifiant des ARN sont aussi importants à caractériser (unité de maturation ; Fig. 1.1 B p. 35). La structure de ce type d'opéron respecte généralement l'ordre ARNr 16S, ARNr 23S et ARNr 5S. En revanche, les ARNt varient tant par le nombre (deux ou trois) que par l'ordre (16S–ARNt–23S–5S–ARNt ou 16S–ARNt–ARNt–23S–5S). Le transcrit primaire (ARN précurseur) de ce type d'opéron subit une étape de maturation complexe pour générer les ARNr et ARNt (unité fonctionnelle).

Parmi les gènes d'ARN, les plus abondants sont les ARN ribosomiques (ARNr) et les ARN de transfert (ARNt ; Tab. 1.1 A p. 37). Au cours de l'expression des gènes de protéines, les ARNr et ARNt qui sont hautement exprimés, interviennent dans l'étape de traduction de l'ARNm. Les

A

Genomic object	Attribute	*B. subtilis*	*E. coli*	*M. tuberculosis*
Chromosome	length (pb)	4214630	4639221	4411532
	G+C %	43,51	50,79	65,61
16S 23S 5S rRNA	number	10	7	1
	chromosome coverage %	1,09	0,70	0,11
tRNA	number	86	82	45
	chromosome coverage %	0,16	0,15	0,08
	length mean (bp)	78,03	86,89	75,22
CDS	number	4107	4274	3996
	chromosome coverage %	87,27	87,21	91,25
	length mean (bp)	895,59	946,63	1007,38
	number (length < 63 pb)	0	5	0
	mutation number	3	251	27
IS	number	0	42	17
Other RNA	number	3	19	2
	chromosome coverage %	0,03	0,05	0,02
Gene	number	4226	4399	4046
	leading %	74,70	54,44	59,19
	lagging %	25,30	45,56	40,81
Intergenic	chromosome coverage %	11,45	11,89	8,54

B

gènes	catalytiques	petits ARN
ssrA	Dégradation de protéines produites par des ARN partiels	ARN 10Sa tmRNA
rnpB	Composant de la ribonucléase P1 qui intervient dans la maturation des ARNt (clivage de l'extrémité 5')	ARN M1

gènes	régulateurs	petits ARN
ssrS	Interaction avec sigma70 (*rpoD*) de l'ARN polymérase lors de la transition de la phase exponentielle à la phase stationnaire	ARN 6S
oxyS	ARN antisens qui empêche la traduction de sigma38 (*rpoS*) dans la réponse au stress oxydatif	
dsrA	ARN antisens qui augmente la traduction de sigma38 (*rpoS*) à basses températures	
micF	ARN antisens qui empêche la traduction de la porine (*ompF*) dans la réponse aux stress oxydatif	

TAB. 1.1 – Caractéristiques des objets génomiques

A) Caractéristiques générales des trois génomes modèles (pour les versions des jeux de réannotation voir p. 213 [Moszer *et al.*, 2002, Rudd, 2000, Camus *et al.*, 2002])

B) Caractéristiques fonctionnelles des petits ARN stables (*small stable RNA*)

ARNr s'associent aux protéines ribosomiques pour former les ribosomes. Les ARN 16S (1500 pb ; gène *rrs* chez *E. coli* K-12) et 5S (120 pb ; *rrf*) sont des molécules de structure du ribosome. L'ARNr 23S (3000 pb ; *rrl*) est l'agent catalytique dans la synthèse de protéines. La sous-unité 30S d'un ribosome contient l'ARN 16S et la sous-unité 50S contient l'ARNr 5S et 23S. Les ARNt (longueur moyenne 80 pb ; Tab. 1.1 A p. 37) sont des adaptateurs dans la synthèse de protéines. Un ARNt reconnaît les codons de l'ARNm et se lie covalemment à l'acide aminé approprié. Il n'y a pas d'équivalence 1-1 entre les codons et les ARNt : les codons d'un acide aminé peuvent être reconnus par plusieurs ARNt, et un ARNt peut reconnaître différents codons d'un même acide aminé. C'est la propriété de flottement (*wobble*) au niveau de la première base de l'anticodon, qui autorise un ARNt à reconnaître plusieurs codons synonymes (qui diffèrent par leur troisième base). Par ailleurs, les gènes d'ARNt peuvent exister en plusieurs copies dans le génome. Par exemple, dans le cas de *M. tuberculosis* H37Rv, 45 ARNt ont été prédits et il existe 3 ARNt qui reconnaissent la méthionine, qui n'est codée que par un seul codon. Au cours du processus d'annotation, on s'attend à trouver au moins un représentant pour chaque acide aminé, donc au minimum 20 ARNt différents. Il y en a évidemment plus, associés aux différents codons synonymes, mais il y a rarement les 61 ARNt correspondant aux 61 codons sens du code génétique dans un génome.

Il existe aussi de petits ARN stables (*small stable RNA (ssr)* ; Tab. 1.1 B p. 37) comme des ARN intervenant dans des processus de régulation ou des ARN catalytiques, composant des enzymes. Par exemple, l'ARN codé par le gène *ssrA* permet l'étiquetage spécifique des protéines dont la biosynthèse est arrêtée au cours de la traduction. On l'appelle aussi ARNtm car il possède les propriétés mixtes d'un ARNt et d'un ARNm [Zwieb *et al.*, 1998]. A l'entrée en phase stationnaire, il se produit des changements profonds de l'expression génétique qui sont encore mal compris : K. Wassarman et G. Storz ont montré que l'ARN codé par *ssrS* est impliqué dans la répression de l'expression à partir d'un promoteur dépendant de sigma70 [Wassarman & Storz, 2000]. Un facteur sigma associé à l'holoenzyme (cœur) donne la spécificité de l'ARN polymérase pour la reconnaissance de tel ou tel type de promoteur (sigma70 est le facteur spécifique des promoteurs exprimés en phase exponentielle). Pourquoi les petits ARN ont-ils été recrutés comme régulateurs des réponses au stress ? Une hypothèse séduisante serait qu'une assez faible quantité d'énergie et un temps court sont nécessaires pour synthétiser ces petits ARN ; ils seraient donc les régulateurs idéaux pour des réponses rapides face aux changements des conditions environnementales.

Un dernier type de gène d'ARN bactérien est celui des introns auto-catalytiques de groupe II, présent dans les ARNm. Ce sont de larges ribozymes capables de s'auto-épisser. Comme les petits ARN sont de petites tailles et que les introns de groupe II sont peu nombreux, ils sont souvent mal annotés dans les génomes bactériens complets. Combien de gènes de petits ARN reste-t-il encore à identifier ? Ils sont difficiles à mettre en évidence par des cribles mutationnels, difficiles aussi à détecter par des tests biochimiques et difficiles à prédire par une analyse informatique. Aussi le catalogue des gènes codant pour des espèces fonctionnelles d'ARN est vraisemblablement très incomplet aujourd'hui.

Pour les trois génomes modèles utilisés dans le cadre de cette thèse, *B. subtilis* (G+C pauvre, *e.g.*

%G+C < 45 [Kunst *et al.*, 1997]), *E. coli* K-12 (G+C moyen, *e.g.* $45 \leq$ %G+C< 55 [Blattner *et al.*, 1997]) et *M. tuberculosis* H37Rv (G+C riche, *e.g.* %G+C ≥ 55 [Cole *et al.*, 1998]), nous disposons d'un jeu de réannotation (ou jeu de référence ; voir p. 213 [Moszer *et al.*, 2002, Rudd, 2000, Camus *et al.*, 2002]). Ils font partie des « gros » génomes puisque la longueur de leur chromosome est supérieure à 4 Mb (Tab. 1.1 A p. 37). Si on compare les annotations de ces génomes, on constate que le nombre de gènes d'ARN chez *M. tuberculosis* H37Rv est inférieur à celui de *B. subtilis* et d'*E. coli* K-12. Les opérons d'ARN existent en un ou plusieurs exemplaires sur le chromosome. Le nombre des copies est grossièrement corrélé à la taille du génome et–ou à la vitesse de croissance des bactéries. Ainsi, il n'existe qu'une copie chez les mycobactéries à croissance lente et chez certains mycoplasmes, deux copies chez les mycobactéries à croissance rapide et chez d'autres mycoplasmes, sept chez les entérobactéries, dix chez *B. subtilis*. La différence de pourcentage de codant et de longueur des CDS pourrait s'expliquer par une différence dans le choix du codon d'initiation entre les différents groupes de (ré)annotation (artefact d'annotation ?). Le génome de *B. subtilis* présente une proportion de gènes transcrits sur le brin précoce beaucoup plus importante que chez les deux autres génomes. Enfin, aucune séquence d'insertion (IS) n'a été répertoriée dans ce dernier génome [Danchin & Sekowska, 1993]. Ce résultat est surprenant puisque 120 IS ont été détectées dans le génome de *Bacillus halodurans*, un autre bacille à Gram positif [Takami *et al.*, 2001].

1.2 Code génétique : usage des codons synonymes et des acides aminés

1.2.1 Distribution des bases et des acides aminés

Les séquences biologiques ne sont pas générées aléatoirement : elles sont riches de redondances et de biais statistiques car elles sont soumises aux processus d'évolution (mutation et sélection naturelle). Lorsque nous parlerons de biais dans les séquences, nous ferons le plus souvent référence à des biais de composition ou de distribution en oligonucléotides, bien qu'il existe d'autres biais, par exemple le biais de distribution du nombre de gènes. Aussi, le biais fait-il référence à un choix différentiel significatif dans l'utilisation d'oligonucléotides particuliers.

Composition du génome en oligonucléotides

La règle de parité 1 (PR1) énonce l'équivalence $A = T$ et $G = C$ dans la molécule d'ADN *double brin* [Chargaff, 1950]. La règle de parité 2 de Chargaff (PR2) énonce l'équivalence $A \sim T$ et $G \sim C$ dans une molécule d'ADN *simple brin* suffisamment longue. Cette seconde règle (conséquence de la première) ne peut être comprise que dans le contexte de l'évolution moléculaire. Quand mutation et sélection affectent symétriquement les deux brins, la matrice de substitutions est symétrique ($P(G \to T) = P(T \to G)$) et la parité est garantie [Sueoka, 1993]. En revanche, s'il existe un biais mutationnel, la règle PR2 n'est plus respectée (voir p. 42).

La première mesure effectuée sur un nouveau génome complet est la composition moyenne des

lettres G et C le long du chromosome simple brin. On calcule tout simplement le pourcentage en G+C du chromosome simple brin ($(N_G + N_C)/N$, avec N le nombre total de nucléotides). Nous avons vu que *B. subtilis* est qualifié de génome G+C pauvre (43,5% G+C, soit 56,5% A+T), *E. coli* K-12 de G+C moyen (50,8% G+C) et *M. tuberculosis* H37Rv de G+C riche (65,6% G+C). La règle PR2 n'est pas respectée car il existe un contexte de mutagenèse asymétrique sur les deux brins (voir p. 42). Chez les procaryotes, la fourchette de variation du pourcentage en G+C entre les génomes varie de 25 à 70%. Nous verrons que les génomes riches en G+C sont particulièrement difficiles à annoter (voir p. 197).

Une seconde mesure permet d'analyser l'hétérogénéité de composition locale de G+C le long du chromosome simple brin au moyen d'une fenêtre glissante dont on doit définir la taille (1000 pb) et le pas (200 pb). On visualise la courbe du pourcentage en G+C le long du génome. Le pourcentage en G+C intra-génomique des procaryotes varie (fourchette de variation de 27 à 67% chez *E. coli* K-12). Cette représentation donne une première idée du nombre et de la taille des régions atypiques, par exemple riches en A+T (caractéristiques de transferts horizontaux ; voir p. 46 et p. 197). Enfin, on étudie la déviation à la parité PR2 par la méthode graphique du G+C-skew[4], qui mesure $(G - C)/(G + C)$ (voir p. 122) et permet l'identification d'asymétries dans les séquences. Les asymétries mettent généralement en évidence les positions de l'origine et du terminus de réplication (voir p. 42).

Le calcul des fréquences en lettres le long de la séquence génomique simple brin peut être généralisé au calcul des fréquences d'un m-uplet le long du chromosome, ceci afin d'étudier l'apparition exceptionnelle de certains mots. En général, chez les organismes procaryotes, les sites reconnus par les enzymes de restriction, souvent palindromiques, GTAC, sont sous-représentés, alors que les sites chi[5] ont tendance à être sur-représentés [Nuel, 2001].

Fréquence de m-uplets dans les CDS

Un gène (ou un ensemble de gènes) peut être défini par un grand nombre de caractères comme son contenu en G+C, en codons ou en acides aminés. Par exemple, calculer les fréquences absolues ou relatives des triplets, en respectant la phase de lecture des CDS, permet d'étudier l'usage du code génétique (Tab. 1.2 p. 41). Il existe en effet $4^3 = 64 - 3$ *codons_stops*, soit 61 *codons* codant 20 acides aminés. Le code génétique est dit dégénéré puisqu'un acide aminé peut être codé par plusieurs codons (on parlera alors de codons synonymes). La distribution des codons doit nécessairement suivre celle des acides aminés qui leurs correspondent dans le code génétique. Par exemple, chez *E. coli* K-12, puisque les protéines contiennent en moyenne 1,5% de tryptophane, on s'attend à trouver 1,5% TGG à l'intérieur des CDS (un seul codon code l'acide aminé tryptophane ; Tab. 1.2 p. 41). A l'exception du tryptophane, de la méthionine (un seul codon), et de l'isoleucine (trois codons), les autres acides aminés sont codés par des duets, des quartets ou des sextets (Tab. 1.2

[4]Le GC-skew est égal à zéro si la parité est respectée.

[5]Le motif chi pour *crossover hotspot initiator* intervient, comme son nom l'indique, dans le processus de recombinaison mais a également une fonction de protection du génome contre l'activité de dégradation des nucléases

p. 41). Le choix du nucléotide en troisième position (et, dans une moindre mesure, en deuxième position) des triplets est moins contrainte par la séquence protéique que celui de la première position. Le duet de la cystéine est rare. La leucine est un acide aminé abondant codé par un sextet, et dont le codon préféré est CTG (Tab. 1.2 p. 41). Etudier les fréquences d'apparition des codons, triplets, 3-uplets[6], permet de mettre en évidence des codons rares et des codons préférés (ou optimaux), un choix différentiel sur les codons, appelé *biais dans l'usage des codons* ou encore *biais dans l'usage du code génétique*. Le biais observé est fonction de l'espèce. On peut étendre ce concept en cherchant à mettre en évidence d'autres biais, par exemple les fréquences d'apparition des di-codons (hexamères ou 6-uplets) dans la phase de lecture des CDS. En effet, lorsque plusieurs codons de probabilités sensiblement équivalentes sont possibles, le choix entre ceux-ci peut être influencé par le voisinage du codon, c'est-à-dire les nucléotides immédiatement en 5′ et/ou en 3′ [Dardel & Képès, 2002].

A

Phe	3,89	His	2,26
Leu	**10,68**	Gln	4,44
Ile	6,02	Asn	3,95
Met	2,78	Lys	4,39
Val	7,11	Asp	5,16
Ser	5,80	Glu	5,76
Pro	4,43	**Cys**	**1,16**
Thr	5,40	Trp	1,53
Ala	9,52	Arg	5,48
Tyr	2,85	Gly	7,39

Hydrophile Hydrophobe Amphipatique

B

Phe	UUU	2,23	Ser	UCU	0,84	Tyr	UAU	1,62	Cys	UGU	0,51
	UUC	1,65		UCC	0,87		UAC	1,22		UGC	0,64
Leu	UUA	1,39		UCA	0,71	TER	UAA	0,20	TER	UGA	0,09
	UUG	1,37		UCG	0,89		UAG	0,02	Trp	UGG	1,52
	CUU	1,10	Pro	CCU	0,70	His	CAU	1,29	Arg	CGU	2,09
	CUC	1,11		CCC	0,54		CAC	0,97		CGC	2,20
	CUA	0,39		CCA	0,84	Gln	CAA	1,53		CGA	0,35
	CUG	**5,30**		CCG	2,33		CAG	2,89		CGG	0,53
Ile	AUU	3,05	Thr	ACU	0,89	Asn	AAU	1,77	Ser	AGU	0,87
	AUC	2,52		ACC	2,35		AAC	2,17		AGC	1,60
	AUA	0,42		ACA	0,70	Lys	AAA	3,36	Arg	AGA	0,20
Met	AUG	2,77		ACG	1,44		AAG	1,01		**AGG**	**0,11**
Val	GUU	1,83	Ala	GCU	1,53	Asp	GAU	3,23	Gly	GGU	2,49
	GUC	1,53		GCC	2,55		GAC	1,92		GGC	2,98
	GUA	1,09		GCA	2,02	Glu	GAA	3,96		GGA	0,79
	GUG	2,63		GCG	3,39		GAG	1,77		GGG	1,11

Tab. 1.2 – Composition moyenne des protéines et usage du code génétique chez *E. coli* K-12
1348627 codons pour un total de 4274 CDS EcoGene17 (le code génétique universel est utilisé).
A) Fréquence relative des 20 acides aminés (pourcentage).
B) Fréquence relative des 64 codons (pourcentage). Les valeurs extrêmes sont indiquées en gras. Les couleurs indiquent la nature des acides aminés selon l'échelle de Kyte et Doolittle (1982).

[6]Un m-uple ou un h-mot désigne une liste ordonnée de toutes les séquences de h bases consécutives sur la séquence. Le 3-uplet(AAA) est la première séquence du 3-uple.

Le biais observé dans l'usage des codons est en fait composé de deux biais : celui en acides aminés et celui en codons synonymes (voir p. 89). Dans le cas d'études portant sur les variations de l'usage des codons synonymes, il est important de se débarrasser du biais en acides aminés (certains étant rares comme la cystéine et d'autres abondants comme la leucine). Au lieu de calculer la fréquence relative des codons, on calcule alors l'usage relatif des codons synonymes (*Relative Synonymous Codon Usage (RSCU)* ; voir p. 122).

Composition des protéines en acides aminés

De manière surprenante, les séquences protéiques sont de composition moyenne relativement constante pour tout les organismes vivants. Il existe certes des protéines membranaires riches en acides aminés hydrophobes, des protéines impliquées dans l'interaction entre protéine et ADN riches en acides aminés basiques (histidine), mais ce sont des tendances générales, indépendantes de l'espèce. Les acides aminés rares sont les acides aminés soufrés, *i.e.* cystéine (composition moyenne du protéome d'*E. coli* K-12 1,16% Tab. 1.2 p. 41 et de la banque complète Swiss-Prot[7] version 42.0 [Apweiler *et al.*, 2004] 1,57%), la méthionine[8] (*E. coli* K-12 2,78% et Swiss-Prot 2,37%), les acides aminés aromatiques, *i.e.* tryptophane (*E. coli* K-12 1,53% et Swiss-Prot 1,18%), la tyrosine (*E. coli* K-12 2,85% et Swiss-Prot 3,11%), et la phénylalanine (*E. coli* K-12 3,89% et Swiss-Prot 4,05%). Les acides aminés abondants correspondent à certains acides aminés hydrophobes : leucine (*E. coli* K-12 10,68% et Swiss-Prot 9,60%), alanine (*E. coli* K-12 9,52% et Swiss-Prot 7,76%) et valine (*E. coli* K-12 7,11% et Swiss-Prot 6,67%), et à certains acides aminés amphipatiques : glycine (*E. coli* K-12 7,39% et Swiss-Prot 6,90%) et sérine (*E. coli* K-12 5,80% et Swiss-Prot 6,94%). Ainsi les codons synonymes du code génétique permettent, à partir de biais de composition en G+C très différents entre les génomes, de coder des polypeptides de biais de composition en acides aminés relativement constants à travers le Vivant.

1.2.2 Origines des biais observés dans les séquences biologiques

La plupart des méthodes d'analyse de séquences reposent sur l'hypothèse qu'il existe des biais de composition en oligonucléotides dans les séquences *naturelles*. De fait, les trois mécanismes fondamentaux que sont la réplication, la transcription et la traduction induisent au sein de la molécule d'ADN des biais de différente nature.

Biais de réplication et de transcription

Les processus d'initiation et de terminaison de la réplication sont régulés par des signaux qui occupent une petite fraction du génome. Différents facteurs peuvent être à l'origine des biais de

[7] `http://us.expasy.org/sprot/relnotes/relstat.html`

[8] La composition des protéines est aussi influencée par le nombre de codons correspondant à chaque acide aminé (alors la méthionine n'est en fait pas rare).

réplication. C'est le cas des contraintes liées à la structure de l'ADN (courbure, super-hélicité) et/ou à la machinerie de la réplication et de la réparation de l'ADN.

Par ailleurs, les mesures du *GC-skew* et de l'*AT-skew* ont permis de mettre en évidence des biais de brin réplicatif[9] sur la séquence chromosomique [Frank & Lobry, 1999].

Selon les génomes, ce calcul reflète deux biais qui s'additionnent ou se soustraient. Le premier biais correspond au biais mutationnel ou biais de composition nucléotidique causé par les mutations. La théorie la plus acceptée pour expliquer ce biais, est l'hypothèse de la désamination de la cytosine [Frank & Lobry, 1999]. Le mécanisme de réplication implique que le brin matrice du brin tardif passe plus de temps dans l'état simple brin que le brin matrice du brin précoce. L'ADN simple brin est bien plus vulnérable aux mutations que l'ADN double brin, la mutation la plus fréquente étant la réaction de désamination hydrolytique de la 5-méthyl-cytosine. La désamination de C vers T et celle moins fréquente de A vers G, dans la matrice du brin tardif, augmente la fréquence de T relativement à celle de A et la fréquence de G relativement à celle de C, dans le brin précoce.

Finalement, l'effet conjoint de ces deux mutations aboutit à un brin précoce biaisé vers G et T contre A et C dans le brin tardif. On comprend donc pourquoi les asymétries identifiées par le *GC-skew* sont plus fortes en troisième position des codons et dans les régions non-codantes (le biais mutationnel n'est pas contre-sélectionné). Le second biais correspond au biais dans la distribution des gènes entre le brin précoce et le brin tardif (*i.e.* chez *B. subtilis* 75% des gènes sont sur le brin précoce). Les régions codantes sont généralement G+A riches (et les régions non-codantes sont A+T riches [Nicolas *et al.*, 2002]) ; le biais de distribution des gènes sur le brin précoce va s'additionner à celui du biais mutationnel qui est responsable d'une plus grande richesse en nucléotides G. Pour s'affranchir du biais de gènes, on calcule le *delta-GC-skew* c'est-à-dire la différence entre le *GC-skew* des gènes du brin tardif et celui des gènes du brin précoce. E. Rocha a montré récemment que le positionnement préférentiel des gènes sur le brin précoce chez *E. coli* K-12 et *B. subtilis* est lié au caractère *essentiel* de ces gènes et non à une forte expression [Rocha & Danchin, 2003]. En effet, durant la réplication du brin tardif, il peut y avoir des collisions frontales entre une ARN polymérase et l'ADN polymérase conduisant à l'interruption de la transcription. Le transcrit avorté peut être traduit en un polypeptide tronqué. Si ce polypeptide appartient à un complexe protéique (ribosome, ARN polymérase, ADN polymérase), il peut produire un phénotype dominant négatif qui peut être délétère pour la bactérie [Rocha & Danchin, 2003].

Biais de traduction

La traduction est le processus énergétique le plus coûteux chez les bactéries. C'est aussi un processus universel puisque les ARN ribosomiques sont conservés dans le règne du Vivant. Au cours des différentes étapes de la traduction (initiation, élongation et terminaison), des contraintes de différentes natures peuvent être à l'origine des biais observés :

[9] La partie du brin direct comprise entre l'origine en $5'$ et le terminus en $3'$ et la partie du brin direct comprise entre le terminus en $5'$ et l'origine en $3'$, correspondent respectivement au brin précoce et au brin tardif.

- L'*initiation* de la traduction nécessite la présence d'un motif RBS, localisé à une certaine distance du codon initiateur. La partie 5' des gènes révèle un biais de composition important (abondance de A entre les positions -30 et +30), dont le rôle est probablement d'éviter la formation de structure secondaire de l'ARNm qui gênerait la mise en place et la progression du ribosome. L'AUG est le codon d'initiation préféré chez toutes les eubactéries, abondance qui chez *B. subtilis*, ne semble ni corrélée au taux d'expression des gènes, ni dépendante de la richesse en GC du génome [Rocha *et al.*, 1999].
- Au cours du *processus d'élongation*, des pressions liées à la vitesse (abondance relative des ARNt isoaccepteurs dans les cellules), et à la qualité (stabilité de l'interaction entre codon et anti-codon) de la traduction s'exercent conjointement. La recherche de l'ARNt correct pour le codon traduit est l'étape limitante de la réaction d'élongation. L'élongation est donc influencée par le biais d'usage des codons synonymes. Les gènes les plus fortement biaisés correspondent généralement aux gènes fortement exprimés en phase exponentielle de croissance. Ils ont adapté leur usage des codons aux ARNt les plus abondants dans la cellule, ARNt qui produisent des interactions codon/anti-codon stables (et non pas de type *wobble*), assurant ainsi une fidélité optimale de la traduction [Ikemura, 1985, Kunisawa *et al.*, 1998]. Les gènes faiblement exprimés subissent une pression sélective insuffisante pour s'adapter à une composition optimale de codons. Ils sont donc plus sensibles au biais mutationnel. Pour ces gènes, l'usage des codons reflète aussi le contenu en GC du génome (GC3), le biais en di-codon, etc. Par ailleurs, les structures secondaire et tertiaire des protéines influencent le choix des codons. Il existe en effet un biais symétrique aux extrémités des polypeptides : sur-représentation des résidus hydrophiles et sous-représentation des résidus hydrophobes. L'ensemble de ces contraintes se traduit majoritairement par un usage irrégulier des codons synonymes suivant les organismes, les codons optimaux étant préférentiellement des trinucléotides de la forme RNY (R, purine :A,G ; Y, pyrimidine : C,T [Shepherd, 1981]). Le biais d'usage des codons synonymes correspondrait donc à l'équilibre entre le biais de mutation sur le génome tendant vers un usage aléatoire des codons synonymes, et les pressions de sélection favorisant l'usage de codons optimaux pour la traduction. C'est ce qu'on appelle la théorie de *selection-mutation-drift* conciliant la théorie neutraliste et la théorie sélectionniste de l'évolution (voir p. 48 [Bulmer, 1991, Smith & Eyre-Walker, 2001]). Sous cette hypothèse, du point de vue du niveau d'expression des gènes par exemple, l'usage des codons synonymes des gènes fortement exprimés pencherait plutôt du côté des mécanismes de mutation-sélection (biais de traduction), alors que celui des gènes faiblement exprimés refléterait plutôt des mécanismes de mutation-dérive (biais mutationnel).
- La *terminaison de la traduction* est essentielle à la fabrication d'un polypeptide fonctionnel. L'efficacité de cette étape dépend de plusieurs facteurs [Rocha *et al.*, 1999] :

  1. la compétition entre le décodage correct ou non du codon stop par les facteurs RF. RF1 reconnaît UAA et UAG alors que RF2 reconnaît UAA et UGA. Chez *E. coli* K-12 RF1 est en concentration cinq fois moins importante que RF2, ce qui explique que l'ordre de

préférence des codons stops soient UAA, UGA puis UAG

2. le contenu en GC de l'organisme influence le choix des codons UAG et UGA. UAA est plus abondant pour les génomes GC pauvres et UGA pour les génomes GC riches
3. les gènes fortement exprimés en phase exponentielle de croissance montrent une nette préférence pour UAA
4. le contexte du codon stop semble aussi jouer un rôle important puisque les bases situées en aval révèlent un biais de composition. De la même façon que pour l'étape d'initiation, le biais de composition de la fin de l'ARNm pourrait éviter la formation de structures secondaires stables.

1.3 Variabilité génétique

Nous avons vu que :

1. les bactéries se reproduisent « à l'identique »
2. elles ont un temps de génération rapide (*M. tuberculosis* H37Rv a un temps de génération de trois semaines, ce qui est toujours bien plus rapide que l'homme qui met en moyenne vingt cinq ans à se reproduire)
3. elles ont une capacité d'adaptabilité remarquable.

Comment font-elles pour adapter correctement leur métabolisme quand les conditions du milieu deviennent difficiles ? D'une part, les variations phénotypiques[10] affectent le comportement de la bactérie. Lorsqu'elles résultent de l'adaptation à diverses conditions extérieures de l'ensemble d'une population bactérienne ayant le même génotype[11], elles sont réversibles, non transmissibles à la descendance, mais spécifiques (non aléatoire). Leur mécanisme est en relation avec l'activité des gènes qui peut être régulée par des systèmes plus ou moins complexes : induction comme dans l'opéron lactose, répression comme dans l'opéron tryptophane. D'autre part, les variations génotypiques affectent le génome bactérien dans sa séquence nucléotidique.

Les variations génétiques dans leur grande majorité ne permettent pas de s'adapter à un nouveau milieu et certaines sont mêmes létales. Mais une population bactérienne compte des milliards d'individus. Il existe donc une chance d'obtenir une variation génétique qui permette de perpétuer une population sur un milieu défavorable (évolution par hasard et sélection). Les deux principes généraux de variation génétique chez les bactéries sont :

1. les altérations du génome (les mutations et les réarrangements)
2. le transfert de gènes (la transformation, la conjugaison, la transduction, la transposition).

Evidemment, ces deux grands principes sont parfois liés. Par exemple les transposons et les phages peuvent à la fois altérer le génome et permettre le transfert de gènes ou encore la recombinaison peut permettre des réarrangements chromosomiques ou des transferts de gènes.

[10] Le phénotype est l'ensemble des propriétés observables d'une cellule.

[11] Le génotype est l'ensemble des gènes (déterminants génétiques) portés par une cellule.

1.3.1 Altérations génomiques

La mutagenèse spontanée (qui arrive naturellement dans l'environnement) regroupe tous les types de mutations (mutations ponctuelles, réarrangements génomiques) provoquées par n'importe quel type d'événement (erreur de réplication, recombinaison).

Les mutations ponctuelles (*point mutation*) regroupent les substitutions d'une paire de bases ou de paires de bases adjacentes (en tandem), et les décalages du cadre de lecture (*frameshift*). La substitution d'une base regroupe les transitions (changement d'une purine (AG) en une purine ou d'une pyrimidine (CT) en une pyrimidine) et les transversions (changement d'une purine en une pyrimidine ou d'une pyrimidine en une purine). Les substitutions sont entraînées par des erreurs dans la fidélité de la réplication (sélectivité de la polymérase, relecture de la polymérase et réparation des mésappariements) ou par des lésions de l'ADN (site sans base). Ces lésions peuvent apparaître spontanément, à cause d'instabilités de l'ADN (désamination de la cytosine méthylée en uracile) ou être provoquées par des agents mutagènes extérieurs (radiations ionisantes, radicaux libres) ou encore par des réactions avec des intermédiaires ou avec des enzymes présents durant le métabolisme normal (dérivés toxiques de l'oxygène, transférases). Ces lésions sont normalement réparées par des systèmes spécialisés [Den Rupp, 1996]. La plupart des *frameshifts* correspondent à l'insertion ou à la délétion d'une ou deux paires de bases (-GC-) dans une suite de deux bases répétées (polyNN -GCGCGC-). Lorsque la fourchette de réplication est bloquée par une lésion (cassure simple brin), il arrive qu'une base se désapparie dans un polyN autorisant un dérapage (*slippage*) d'un brin par rapport à l'autre, et conduisant alors à une base en plus ou en moins dans le brin nouvellement synthétisé [Hutchinson, 1996].

Les réarrangements chromosomiques reposent sur un mécanisme de recombinaison homologue. Celle-ci est un processus fondamental présent chez tous les organismes et qui permet de réarranger les gènes ou des parties de gènes à l'intérieur d'un chromosome ou entre deux chromosomes, de limiter la divergence de séquences d'ADN répétées et de réparer l'ADN endommagé. Elle implique un échange d'information génétique entre deux séquences similaires. Les réarrangements chromosomiques regroupent des délétions, des duplications en tandem (insertion d'une séquence qui existe déjà dans le génome) et des inversions. Ils peuvent avoir lieu entre deux chromosomes frères issus de la fourche de réplication ou à l'intérieur d'un chromosome [Roth *et al.*, 1996].

1.3.2 Transferts de gènes

Les altérations génomiques que nous venons de décrire ont en commun leur caractère endogène (transmission verticale). A l'inverse, le transfert de gènes entre deux organismes implique la notion de matériel exogène, d'où la notion de transfert horizontal (*horizontal gene transfert*, *HGT*). Quatre processus permettent les transferts de matériel génétique : la transformation (plasmide), la conjugaison (épisome), la transduction (phage) et la transposition (séquence d'insertion ou IS). De façon générale, l'ADN transféré ne peut se perpétuer dans la bactérie réceptrice que s'il est déjà dans un réplicon (plasmide) ou s'il est immédiatement intégré dans un réplicon (chromosome

ou plasmide) par recombinaison (homologue, site-spécifique). La recombinaison site-spécifique se fait entre deux segments d'ADN définis et similaires, reconnues par des enzymes spécifiques qui catalysent la coupure et la religature. Ces enzymes sont des recombinases spécialisées comme les intégrases de phage.

Les *quatre processus* évoqués plus haut peuvent être décrits de la manière suivante [Saunders *et al.*, 1999] :

1. La transformation est le transfert passif d'ADN d'une bactérie donatrice à une bactérie réceptrice, dite en état de compétence. En 1944, suite à l'expérience de F. Griffith de 1928, O. Avery démontre d'une part que l'ADN est le support de l'information génétique et d'autre part que les bactéries sont capables d'intégrer des fragments d'ADN chromosomiques provenant du milieu extérieur [Avery *et al.*, 1944]. En laboratoire, il est plus efficace de transformer les bactéries par des plasmides plutôt que par des fragments d'ADN linéaires, rapidement dégradés par les bactéries.

2. La conjugaison est un transfert d'ADN entre une bactérie donatrice (mâle, ou F+) et une bactérie réceptrice (femelle, ou F-), qui nécessite le contact et l'appariement entre les bactéries, et repose sur la présence dans la bactérie F+ d'un facteur de sexualité ou de fertilité (facteur F). Le facteur F est un plasmide conjugatif, c'est-à-dire qu'un des brins est transféré par son extrémité $5'$ au travers d'un pont cytoplasmique (pilus sexuel) par le mécanisme réplicatif du cercle roulant (*rolling circle*). Le facteur F est un épisome car il a aussi la capacité de s'intégrer dans le chromosome bactérien et de s'en exciser. Lors de l'excision le plasmide F peut emporter un fragment d'ADN chromosomique. Il est alors appelé (F'). Enfin, si la conjugaison a lieu au moment où le plasmide F est intégré dans le chromosome, il est alors possible de transférer tout le chromosome bactérien.

3. La transduction est le transfert d'ADN par l'intermédiaire de certains bactériophages, dits tempérés. En effet les autres bactériophages, dits virulents, ne peuvent transférer de l'ADN puisque après multiplication dans les bactéries, ils les lysent systématiquement. Le phage tempéré peut exister sous sa forme prophage, intégré dans le chromosome d'une bactérie dite lysogène, ou sous sa forme virulente, après induction du cycle lytique. Au moment de son encapsidation et de la lyse bactérienne, il peut incorporer des fragments d'ADN bactériens qui seront injectés dans le cytoplasme bactérien lors de l'infection suivante. Ces fragments bactériens peuvent alors être intégrés par recombinaison dans le chromosome de la nouvelle bactérie infectée. Le phage tempéré a seulement un rôle de vecteur dans le mécanisme de transduction. Il existe aussi, la conversion lysogénique, où le génome du phage tempéré contient un ou plusieurs gènes utiles à la bactérie comme des gènes codant des toxines ou des facteurs antigéniques. Leur expression dans toutes les bactéries est liée à l'état lysogène. Il disparaît avec la perte de celui-ci.

4. Les transposons sont des séquences d'ADN capables de changer de localisation dans le génome sans jamais apparaître à l'état libre. La recombinaison des transposons n'est ni une

recombinaison spécifique, ni une recombinaison homologue. D'une part, la recombinaison des transposons ne nécessite pas deux séquences homologues ou deux sites spécifiques mais une seule séquence cible très courte (TCGAT). Il existe donc de très nombreux sites potentiels d'intégration. D'autre part, elle nécessite une réplication de l'ADN. Dans le cas général des transposons duplicatifs, si la transposition est conservative, elle ne nécessite que la réplication de la séquence cible dupliquée, alors que si elle est réplicative, elle nécessite la réplication du transposon et de la séquence cible dupliquée [Craig, 1996]. Il existe trois types d'éléments transposables : les IS, les transposons composés et les transposons non composés. L'IS1 est constituée de deux gènes, *insA* et *insB*, flanquée de deux séquences répétées inversées (IR ; ACAGTTCAG-*insA*-*insB*-CTGAACTGT). Les deux polypeptides codés par *insA* et *insB*, s'assemblent pour former la transposase. Le transposon composé IS10-Tn10 est constitué du gène de résistance à la tétracycline flanqué par deux IS10. Le transposon non composé Tn3 est constitué par une transposase, un site de recombinaison, une résolvase, un gène de résistance à l'ampicilline, le tout flanqué par deux IR [Craig, 1996].

Enfin, il existe une cinquième façon de transférer de l'ADN : un intégron permet de capturer une cassette et d'en exprimer les gènes [Ochman *et al.*, 2000]. Un intégron est un élément immobile constitué d'une intégrase, d'un promoteur et d'un site-spécifique, et porté par un réplicon. L'intégrase est capable d'insérer ou d'exciser une cassette (élément mobile) par recombinaison site spécifique. Plus de 60 cassettes[12] impliquées dans la résistance aux antibiotiques ou aux antiseptiques ont été décrites.

1.4 Monde procaryote et phylogénie

Le monde vivant est divisé en trois grands domaines : les archaea (archaebactéries) et les bactéries (eubactéries) qui constituent les organismes procaryotes, puis les eucaryotes (Fig. 1.4 p. 55). Les virus ne sont pas des êtres vivants car ils n'ont pas de métabolisme propre [Danchin, 1998], pas de reproduction autonome et pas de membrane plasmique. En avril 2004, parmi les génomes complets publiés, 18 génomes d'archaea, 142 génomes de bactéries et 26 génomes d'eucaryotes sont disponibles (d'après le site GOLD[13]).

1.4.1 Classification des espèces

L'espèce constitue l'unité de base de la classification bactérienne. La définition classique d'une espèce biologique (une communauté d'êtres vivants reconnaissables par leurs caractères et capables de se reproduire sexuellement entre eux en donnant naissance à une progéniture fertile) n'est pas applicable aux procaryotes [Cohan, 2002]. Les bactériologistes ont dû élaborer une définition originale de l'espèce. En bactériologie, une espèce est constituée par sa souche type et par l'ensemble

[12] `http://www.microbes-edu.org/etudiant/gene4.html`
[13] `http://www.genomesonline.org/`

des souches considérées comme suffisamment proches de la souche type pour être incluses au sein de la même espèce. La systématique a pour but de classer les êtres vivants de manière rationnelle en se basant sur les ressemblances et sur les relations qui existent entre eux. Elle repose sur deux disciplines, la taxonomie et la nomenclature. La taxonomie est la science qui permet de classer les organismes en groupes d'affinité ou taxa, et la nomenclature est l'ensemble des règles qui président à l'attribution d'un nom à chaque taxon. Le nom résume l'ensemble des caractéristiques d'un taxon et l'utilisation d'une nomenclature correcte permet à tous les protagonistes d'une même discipline de se comprendre sans ambiguïté. Par exemple, le placement d'une souche bactérienne dans la famille des *Enterobacteriaceae* permet à un bactériologiste de comprendre que cette souche rassemble des bacilles à Gram négatif, non sporulés, aéro-anaérobies, etc. Pour être fonctionnel, un système de nomenclature doit être rigoureux, précis, unique, universel et non ambigu (annexe A p. 383). Il existe différentes approches taxonomiques, dont les trois principales sont les suivantes :

1. La classification phénotypique utilise un faible nombre de caractères considérés comme importants tels que la morphologie, la mise en évidence d'un caractère biochimique jugé essentiel, l'habitat, le pouvoir pathogène, etc. Cette classification a l'inconvénient de ne refléter qu'une quantité d'information réduite. De plus le choix des critères qualifiés d'importants est subjectif et il peut varier d'un auteur à un autre, ce qui est une source potentielle d'instabilité.
2. Il fallut attendre la deuxième moitié du $20^{\text{ème}}$ siècle pour qu'une taxonomie phylogénétique commence à se mettre en place.
   - En 1950, Chargaff [Chargaff, 1950] montre que le contenu en bases puriques et en bases pyrimidiques de l'ADN peut varier d'un individu à un autre mais est constant pour les individus d'une même espèce. Actuellement, on admet que des bactéries dont les G+C diffèrent de plus de 5% ne peuvent appartenir à une même espèce et que des bactéries dont les G+C diffèrent de plus de 10% ne peuvent appartenir à un même genre. Bien sûr, des valeurs du pourcentage en G+C identiques n'impliquent pas que les bactéries sont proches car les séquences peuvent être très différentes.
   - Les méthodes d'hybridation ADN-ADN sont basées sur le fait que deux molécules d'ADN dénaturées peuvent se réassocier à condition de présenter une homologie. La renaturation est réalisée à partir d'un mélange de deux ADN dénaturés provenant de bactéries différentes. Dans ces conditions, on obtient d'autant plus de duplex hétérologues que les séquences d'ADN des micro-organismes sont proches. Pour reconnaître la provenance de chaque brin d'ADN dans les hybrides, on marque l'un des ADN par un isotope radioactif ou par une enzyme. La solidité, et donc la spécificité, des hybrides sont appréciées par la mesure de la stabilité thermique.
   - En 1962, L. Pauling et E. Zuckerkandl ont émis le concept *d'horloge moléculaire* [Kimura, 1983] : la vitesse de l'évolution est constante, les mutations qui surviennent dans le génome n'ont pas nécessairement de conséquences phénotypiques mais elles sont étroitement corrélées avec le temps (théorie neutraliste de l'évolution). Dans ces conditions, il est possible de construire un arbre généalogique (phylogénétique) en utilisant des méthodes

bioinformatiques. Le principe de base consiste à comparer des gènes homologues de fonction homologue (séquences similaires descendant d'un ancêtre commun qui codent des polypeptides ayant conservé une fonction identique au cours du temps). Le choix des séquences à comparer pose un problème car il est difficile de trouver une molécule qui soit présente et homologue chez tous les organismes et qui présente des niveaux successifs d'information. En effet, pour comparer des organismes très éloignés, il faut utiliser des séquences qui restent sensiblement conservées durant des centaines de millions d'années, tandis que la comparaison d'organismes proches requiert l'étude de séquences où des mutations se seront accumulées en quelques millions d'années. Ainsi, les ARNr ont été choisis en taxonomie pour quatre raisons principales :

(a) ils sont présents dans toutes les cellules ce qui permet des comparaisons entre procaryotes et eucaryotes

(b) ils ont une structure bien conservée car toutes modifications pourraient avoir des conséquences importantes sur les synthèses protéiques

(c) il existe des portions d'ARNr dont la séquence est identique chez tous les êtres vivants

(d) ils sont abondants dans la cellule et faciles à purifier.

L'ARN 16S est le plus utilisé car il est plus facile à analyser que l'ARN 23S et plus riche en information que l'ARNr 5S. Ces séquences correspondent à des oligonucléotides présents dans un groupe phylogénique donné mais absent dans la plupart des autres groupes.

3. Les termes de taxonomie mixte et consensuelle (*polyphasic taxonomy*) font référence à une classification qui tient compte d'un maximum de données : données génétiques, données phénotypiques, données chimiotaxonomiques, données écologiques, etc.

La définition d'une espèce bactérienne diffère en fonction de l'approche taxonomique retenue. Selon [Stackebrandt *et al.*, 2002], le Comité *ad hoc* pour la réévaluation de la définition d'espèces en bactériologie, la définition d'une espèce génomique (*genomospecies*) rentre dans le cadre de la taxonomie mixte et consensuelle.

Elle prend en compte des critères de la taxonomie : (1) phylogénétique (hybridation ADN-ADN, ARN 16S et pourcentage en GC) et (2) phénotypique (morphologie, mobilité, biochimie). Autant que possible, la description d'une nouvelle espèce devrait reposer sur plusieurs souches. De plus, la définition d'une nouvelle espèce doit recevoir l'assentiment de nombreux taxonomistes et sa description doit suivre un plan cohérent et standardisé. Pour les bactéries d'intérêt médical ou vétérinaire, l'écologie et/ou le pouvoir pathogène peuvent prendre le pas sur les critères génétiques et conduire à conserver des nomenclatures distinctes pour des taxons très proches sur le plan génétique, par exemple :

– *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium pinnipedii* et *Mycobacterium tuberculosis* sont des variants d'un seul clone qui devrait être appelé *Mycobacterium tuberculosis*, historiquement découvert en premier.

– *Yersinia pestis* devrait être considérée comme une sous-espèce de *Yersinia pseudotuberculosis*, mais la Commission Judiciaire a rejeté l'appellation de *Yersinia pseudotuberculosis subsp. pestis*.

Dans le cadre d'un travail de génomique comparative, il est essentiel d'avoir un point de vue global sur la classification des génomes procaryotes complets. Le site *Microbial Genome Database for Comparative Analysis* (*MBGD*)[14] [Uchiyama, 2003] fournit une taxonomie des génomes procaryotes complets basée sur celle du *National Center for Biotechnology Information* (*NCBI*)[15]. La taxonomie du NCBI regroupe 101.148 espèces (82.202 eucaryotes, 12.254 procaryotes et 6.692 virus). C'est une taxonomie mixte et consensuelle constituée à partir de différentes ressources comme le manuel de systématique bactériologique de Bergey[16], la nomenclature bactérienne DSMZ[17], la liste des noms des taxons bactériens ayant un statut officiel dans la nomenclature[18], le projet de base d'ARNr (*Ribosomal Database Project* (*RDP-II*)[19] [Cole *et al.*, 2003]), etc. La taxonomie mixte et consensuelle des génomes complets de MBGD présente les différents phyla (ou embranchements) dans l'ordre alphabétique et ne permet donc pas de connaître l'histoire de ces phyla. Ainsi, n'ayant pas trouvé d'arbre des génomes procaryotes complets fondé sur la classification mixte et consensuelle, nous nous sommes penchés sur les méthodes de reconstruction d'arbres phylogénétiques. Ceci afin de pouvoir répondre à des questions du type : le phylum des *Actinobactéries* auquel appartient *M. tuberculosis* H37Rv est-il un groupe frère des *Firmicutes* (*B. subtilis*) ou des *Protéobactéries* (*E. coli* K-12) ?

1.4.2 Reconstruction d'arbres phylogénétiques

En phylogénie moléculaire, de nombreuses approches sont possibles car il existe différents :

1. modèles de classification (hiérarchique, pyramidale, réticulaire),
2. méthodes de reconstruction d'arbre (distance, parcimonie et vraisemblance)
3. types de séquences (une ou plusieurs séquences d'ARN ou de polypeptides),
4. types de données à classer (distance pour chaque couple (*pairwise distance*), nombre de mutations en chaque site, proportion de gènes communs *i.e.* contenu en gènes, conservation de l'ordre des gènes, conservation de signatures de séquences comme les indels [Gupta & Griffiths, 2002], ou signatures génomiques [Karlin *et al.*, 2002]).

Néanmoins, ces différentes approches partent généralement d'un alignement de séquences. La recherche de similitudes entre deux séquences à partir des alignements permet de mettre en évidence

[14] `http://mbgd.genome.ad.jp/htbin/create_tax`
[15] `http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html`
[16] `http://www.cme.msu.edu/bergeys/`
[17] `http://www.dsmz.de/bactnom/bactname.htm`
[18] `http://www.bacterio.cict.fr/`
[19] Le site du RDP est très riche en séquences et il n'est pas facile d'extraire celles des génomes complets pour construire des arbres (`http://rdp.cme.msu.edu/html/` version 8.1, 16.227 séquences d'ARN 16S prokaryotes issues de INSD). Parmi, les arbres d'ARN 16S précalculés proposés sur ce site aucun n'est dédié aux génomes complets procaryotes.

une homologie de séquences. Cependant, une similitude significative entre deux séquences n'est pas toujours une preuve de leur homologie, et *a fortiori*, une homologie entre deux séquences n'est pas toujours une preuve d'homologie de fonction. En évolution moléculaire, deux séquences sont dites homologues si elles dérivent d'un ancêtre commun. Il existe trois grandes classes d'homologie de séquences [Fitch, 2000] :

1. les gènes orthologues, qui ont divergé à la suite d'un événement de spéciation (dans ce cas et dans ce cas seulement, l'histoire des gènes se confond avec celle des espèces)
2. deux gènes sont dits paralogues s'ils ont divergé à la suite d'un événement de duplication, par opposition au premier cas
3. on désigne par xénologues des gènes dont l'homologie est expliquée par les transferts horizontaux (ou latéraux) de matériel génétique d'un organisme à un autre.

Deux séquences sont dites analogues, par opposition à homologues, si elles descendent d'ancêtres sans lien de parenté et si leurs similitudes sont dues à une convergence évolutive[20] [Fitch, 2000]. Les différents opérons d'ARNr d'une espèce peuvent être issus de duplications et/ou de transferts horizontaux. Le choix d'un ARN 16S pour représenter l'espèce est en théorie délicat car il s'agit de trouver le gène ancestral parmi des paralogues et/ou des xénologues. En pratique, il est possible de prendre soit tous les ARN 16S d'une espèce, soit d'en choisir un au hasard, après avoir vérifié qu'ils sont tous très similaires. De plus, sachant que de nombreux chromosomes complets sont accessibles, on peut reprocher aux arbres basés sur les séquences d'ARN 16S de ne prendre en compte qu'une faible partie de l'information. En revanche, les méthodes de phylogénie basées sur les génomes complets, comparent les gènes orthologues communs à tous ces génomes, mais plus le nombre de génomes est important et plus le nombre gènes orthologues diminue. Par ailleurs, ce phénomène est accentué si on prend soin d'écarter les gènes paralogues et xénologues.

A l'heure actuelle, on dénombre vingt-six phyla de bactéries cultivables (cinquante-deux au total [Rappe & Giovannoni, 2003]). Des génomes complets sont disponibles pour seulement douze de ces phyla et six d'entre eux ne sont représentés que par une seule espèce : *Aquificae*, *Bacteroidetes/Chlorobi*, *Deinococcus-Thermus*, *Fusobacteria*, *Planctomycetes* et *Thermotogae*. Il reste donc beaucoup de génomes à séquencer si l'on veut représenter le monde bactérien de manière équitable.

Nous avons confronté les résultats de trois approches de reconstruction d'arbres phylogénétiques des génomes procaryotes complets. Deux utilisent les séquences des protéomes : l'une est fondée sur le contenu en gènes, et l'autre sur l'ordre des gènes (serveur *SHOT*[21] [Korbel *et al.*, 2002]). La troisième approche, plus classique, utilise uniquement la séquence de l'ARN 16S (à l'aide du logiciel *MEGA2*[22] [Kumar *et al.*, 2001]). Pour ces trois approches, nous avons choisi l'algorithme de classification hiérarchique ascendante du *Neighbor Joining* (*NJ*), fondée sur un indice de distance (voir p. 136 [Nei & Kumar, 2000c]).

[20] Les ailes des oiseaux et des chauves-souris ont des caractères analogues, mais elles n'ont aucune origine commune.

[21] *Shared Ortholog and Gene Order Tree Reconstruction Tool*, `http://www.bork.embl-heidelberg.de/~korbel/SHOT/`.

[22] *Molecular Evolutionary Genetics Analysis*, `http://www.megasoftware.net/`.

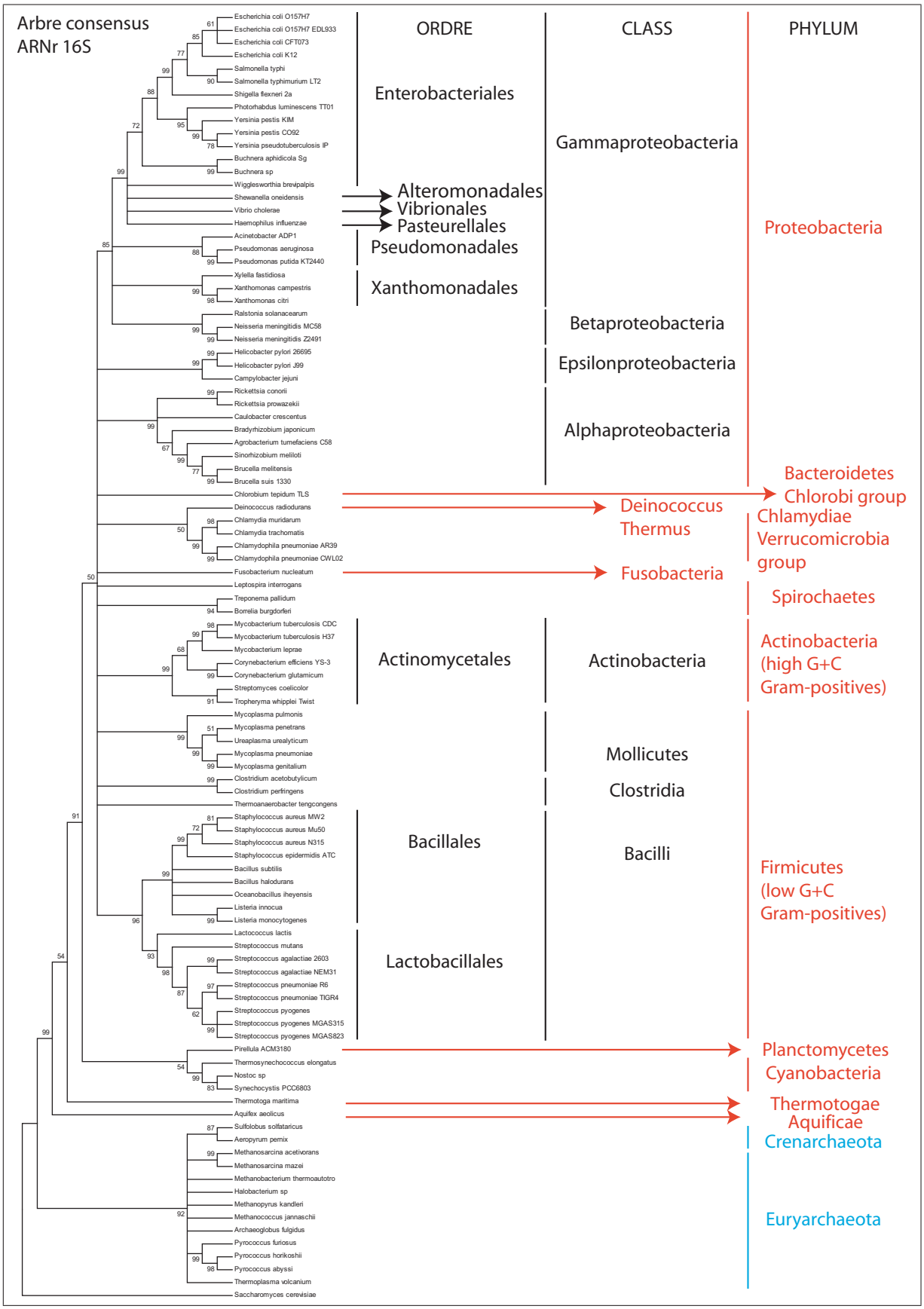


FIG. 1.2 – Arbre phylogénétique d'ARN 16S de génomes procaryotes complets

En ce qui concerne la méthodologie, se référer à l'annexe A.1.1 p. 383. Les classes des *Protéobactéries* Delta et Epsilon appartiennent au sous-phylum Delta-Epsilon.

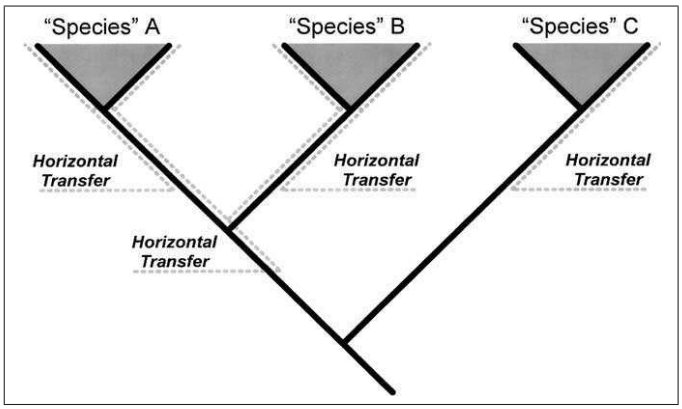


FIG. 1.3 – Liens de parenté parmi les taxons bactériens
En raison des recombinaisons homologues, les phylogénies dérivées à partir de différents gènes ne sont pas congruentes pour une *espèce* bactérienne : ceci est décrit par les triangles gris délimitant la variation génétique à l'intérieur d'un clade. L'ensemble, comprenant un ancêtre et tous ses descendants, définit un taxon monophylétique. En revanche, il existe une barrière qui empêche le mécanisme de recombinaison homologue entre des régions d'ADN non homologues au cours des échanges de gènes et qui est gouvernée par la reconnaissance de mésappariements. De cette absence de recombinaison homologue entre les taxa (ensemble d'organismes contenus dans un niveau hiérarchique d'une classification) résultent des phylogénies congruentes pour les différents gènes. Le transfert horizontal par recombinaisons illégitimes (représenté par des lignes pointillées) peut introduire des gènes dans ces lignées (suite d'ancêtres et de descendants, différentiable des autres lignées). Cependant, tant que les taxa donneurs ne sont pas inclus dans l'analyse phylogénétique, les prédictions de ce modèle restent valides. Ce modèle s'applique donc strictement au cas de phylogénies *locales* (groupe de taxa qui sont proches parents). L'auteur de cette figure est J. Lawrence [Lawrence, 2002].

Les quatre classes du phylum des *Protéobactéries* sont groupées sur l'arbre de l'ordre des gènes alors que les *Protéobactéries* Epsilon forment un groupe à part avec le phylum *Aquificae*, sur l'arbre du contenu en gènes (FIG. A.1 p. 384 et FIG. A.2 p. 385 en annexe). Sur l'arbre de l'ARN16S, seules les classes des *Protéobactéries* Gamma (*E. coli* K-12) et des *Protéobactéries* Beta sont groupées (FIG. 1.2 p. 53). Ces résultats suggèrent que la classe des *Protébactéries delta-epsilon* serait la première du phylum à avoir émergé. Chez les *Firmicutes*, les deux classes des *Mollicutes* et des *Bacilli* (*B. subtilis*) sont groupées sur les arbres du contenu en gènes et de l'ordre des gènes, alors qu'elles sont séparées sur l'arbre de l'ARN16S. Sur l'arbre du contenu en gènes, les *Actinobactéries* comme *M. tuberculosis* H37Rv sont groupés avec les *Protéobactéries*, alors que sur l'arbre de l'ordre des gènes, ce phylum est le groupe frère des *Firmicutes*. Pour les *Actinobactéries* il n'y a pas de congruence entre les trois classifications.

Ainsi, si on observe une congruence entre des arbres obtenus par différentes méthodes de reconstruction phylogénétique, alors la topologie obtenue est la plus raisonnable. Mais elle ne reflète cependant pas toujours la réalité. Les trois arbres, ARN 16S, contenu en gènes et ordre des gènes (FIG. 1.2 p. 53, FIG. A.1 p. 384 et FIG. A.2 p. 385 en annexe), sont globalement semblables mais présentent un certain nombre de différences. S'il existe un arbre universel de la vie, il devrait refléter les relations d'évolution d'organismes complets, et non pas seulement de simples gènes. Cependant, de nombreuses questions concernant la phylogénie des génomes complets procaryotes doivent être élucidées.

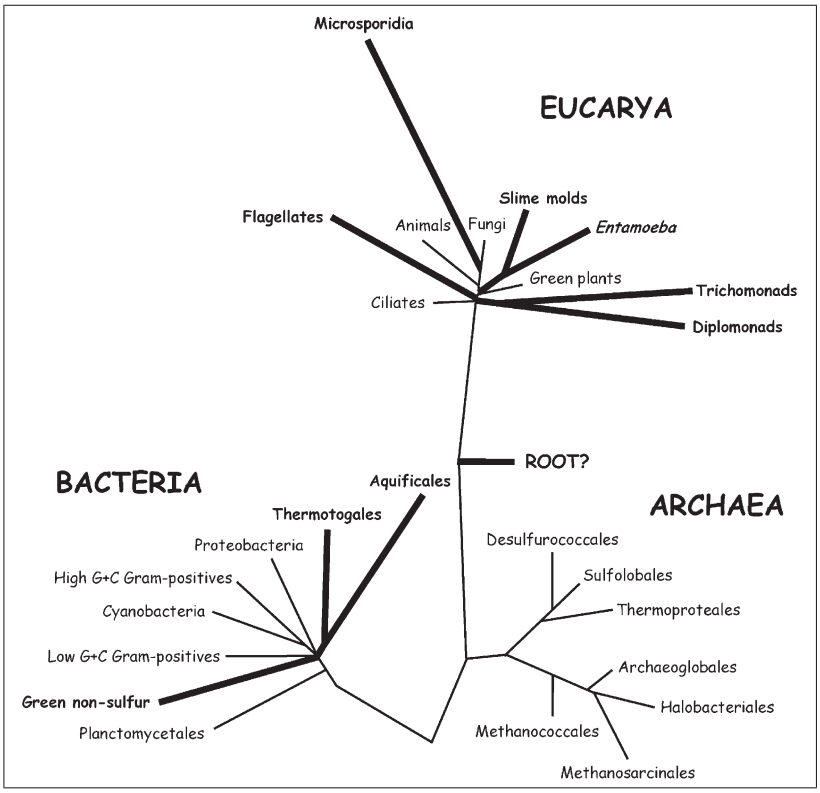


FIG. 1.4 – Une vision revisitée de l'arbre universel de la vie

Les lignes épaisses correspondent à des branches qui peuvent être affectées par des phénomènes d'attraction des longues branches (le placement de la racine, l'émergence du premier taxon bactérien). Elles ont été repositionnées dans la phylogénie soit par une nouvelle analyse des ARNr (l'émergence du premier taxon bactérien) ou de gènes anciennement dupliqués (la racine) soit par l'analyse de nouveaux marqueurs (groupement des *Slime molds* (*Dictyostelium*) avec les *Entamoeba* qui sont aussi des amibes). Certaines révisions sont robustes (les *Thermotogales* avec les *Aquificales*) tandis que d'autres restent ouvertes (la racine représentée par le dernier ancêtre commun universel *Last Universel Common Ancestor* (*LUCA*) »). Les auteurs de cette figure sont S. Gribaldo et H. Philippe [Gribaldo & Philippe, 2002].

R. Jain et J. Lawrence distinguent, dans le transfert horizontal de gènes chez les procaryotes, le transfert intra-spécifique (entre deux individus d'une même espèce) du transfert inter-spécifique (entre deux individus d'espèces différentes) [Jain *et al.*, 2002, Lawrence, 2002]. Pour J. Lawrence il s'agit d'*échanges de gènes* par recombinaisons homologues à l'intérieur d'une espèce et de *transfert de gènes* par recombinaisons illégitimes (site-spécifique ou par transposition) entre espèces (Fig. 1.3 p. 54). Quelle est la part de transmission verticale (altérations du génome) et de transmission horizontale (transferts de gènes) dans les mécanismes d'évolution des procaryotes ? Comment détecter le transfert horizontal en particulier quand c'est un transfert ancien entre deux espèces ou quand il s'est produit à l'intérieur d'une même espèce ? De manière générale, les problèmes d'incongruence entre des phylogénies basés sur des approches différentes sont liés à deux problèmes principaux. D'une part, les phylogénies sont faussées si les méthodes sont appliquées à des gènes xénologues ou paralogues, non identifiés, au lieu de gènes orthologues comme il se devrait ; d'autre part, les méthodes mathématiques ne prennent pas en compte certaines caractéristiques biologiques comme le biais de composition, les vitesses d'évolution différentes entre les espèces (attraction des longues branches), etc. [Gribaldo & Philippe, 2002].

On comprend donc la difficulté à trouver des consensus. En particulier chez les bactéries, il existe un débat pour savoir quel est le phylum bactérien qui a émergé le plus tôt. Par exemple, l'arbre de la figure 1.2 p. 53, en accord avec la topologie classique de l'arbre universel de la vie, fait émerger le *phylum Aquificae* (hyperthermophile) en premier. Mais d'après S. Gribaldo il s'agit d'un artefact ; c'est le phylum des *Planctomycétales* qui aurait émergé en premier (Fig. 1.4 p. 55 ; [Gribaldo & Philippe, 2002]). Les caractéristiques les plus intrigantes des *Planctomycétales* sont l'absence de peptidoglycane dans leur paroi et la présence d'une membrane autour du chromosome délimitant le riboplasme (compartimentation complexe). Si l'émergence primitive des planctomycétales est confirmée, l'origine des bactéries doit être sérieusement reconsidérée (l'ancêtre commun aux bactéries possédait-il une compartimentation complexe ? Y a-t-il eu convergence évolutive du caractère hyperthermophile chez les *archaea* et chez les bactéries ?). Enfin, J. P. Gogarten, W. F. Doolittle et J. G. Lawrence, en s'appuyant sur les évidences de transferts horizontaux chez les procaryotes [Martin, 1999], remettent en cause notre compréhension de l'évolution des procaryotes : nous cherchons l'ancêtre commun des espèces mais une hypothèse plus réaliste serait de s'intéresser simplement aux ancêtres communs des gènes [Gogarten *et al.*, 2002].

1.5 Raisons d'explorer les génomes et protéomes bactériens

Avant de clore cette section, il nous a semblé important de rappeler quelques propriétés qui en font des modèles privilégiés pour l'étude des organismes vivants, tant d'un point de vue fondamental que d'un point de vue appliqué. Les perspectives de recherche sont énormes puisqu'on estime que moins d'1% du monde procaryote est connu [Amann *et al.*, 1995]. La compacité de leur génome, la rapidité de leur croissance, leurs structures opéroniques, leurs mécanismes d'évolution diversifiés (chromosomes mosaïques), etc., confèrent à ces êtres unicellulaires, une grande plasticité adaptative

(*fitness*).

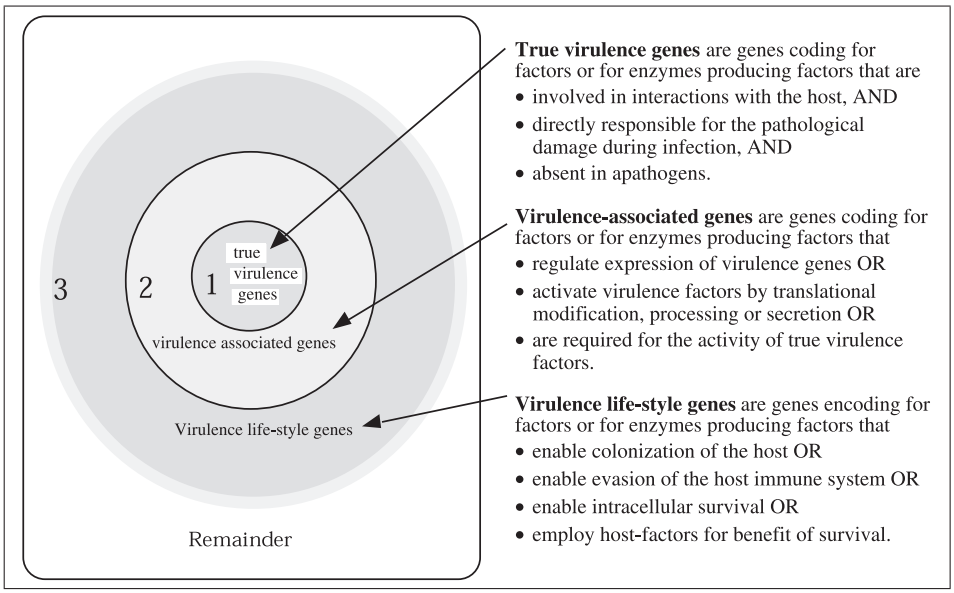


FIG. 1.5 – Définition de la virulence

Le nombre de gènes appelés gènes de virulence (gènes codant des fonctions de virulence) dépend de l'ontologie du concept de virulence. Le nombre de gènes de virulence et de gènes associés à la virulence sont inclus dans des cercles concentriques. Dans la collection 1, seuls les facteur de virulence qui sont directement impliqués dans les causes de la maladie sont inclus (*gènes de « vraie » virulence*). L'addition des *gènes associés à la virulence* augmente le nombre de gènes de virulence identifiés et donc la taille du deuxième cercle. Le pool de gènes identifiés par inactivation et par caractérisation phénotypique inclut tous les gènes qui mènent à un phénotype de virulence atténuée comme les gènes caractéristiques du style de vie des pathogènes. Les auteurs de cette figure sont T. Wassenaar et W. Gaastra [Wassenaar & Gaastra, 2001].

Dans le domaine des biotechnologies, les applications relatives à l'analyse des génomes bactériens sont aussi multiples : certaines bactéries comme *E. coli* K-12 sont devenues de véritables instruments de production (amplification des banques d'ADN pour le séquençage, production d'acides aminés essentiels pour l'agroalimentaire).

En médecine, les bactéries sont étudiées du point de vue de leur pathogénie à travers, par exemple, l'identification d'îlots de pathogénie et la détection de gènes de résistance aux antibiotiques et aux désinfectants. Par ailleurs, l'étude de la virulence (mesure qualitative de la pathogénie) et de l'agressivité (mesure quantitative de la pathogénie) des bactéries pathogènes nécessite d'étudier sa relation avec l'hôte qui exprime la maladie. Les nombreux mécanismes de virulence mettent en jeu des gènes d'importance variable : gènes de « vraie » virulence (codant une toxine), gènes associés à la virulence (codant un activateur de l'expression de la toxine) et gènes caractéristiques du style de vie de la bactérie pathogène dans l'hôte (FIG. 1.5 p. 57 [Wassenaar & Gaastra, 2001]). Dans cette dernière catégorie, on peut citer par exemple des gènes codant des antigènes pour les bactéries nichant dans les voies respiratoires, des gènes codant des invasines pour les bactéries intracellulaires, des gènes codant des adhésines pour les bactéries entériques, etc. L'analyse des génomes de bactéries pathogènes fournit une vision inédite de leur virulence, de leur évolution et de leur adaptation à leur hôte. Ces résultats d'analyse ont permis de mettre en évidence différentes stratégies de

virulence en fonction des groupes de bactéries pathogènes comme les transferts latéraux de gènes chez les bactéries entériques (les îlots de pathogénie chez *Y. pestis* CO92), la perte de fonction chez les pathogènes obligatoires intracellulaires (les pseudogènes chez *Rickettsia prowazekii*) et les variations antigéniques chez les pathogènes des muqueuses (les répétitions *Neisseria meningitidis* Z2491 (Serogroup A)) [Wren, 2000].

Ainsi les procaryotes, sous leurs apparences rudimentaires, révèlent une diversité inattendue de réseaux cellulaires (cascades de régulations et voies métaboliques) participant activement à l'équilibre complexe du monde vivant. Les projets des centres de séquençage ont d'abord été orientés vers les bactéries pathogènes et les bactéries utilisées en biotechnologie, alors que les nouveaux projets sont plus axés sur les bactéries et les *archaea* de l'environnement [Amann, 2002].

Chapitre 2

Ressources disponibles pour l'étude des génomes procaryotes

Quel que soit le domaine d'étude, il ne suffit pas de produire de l'information, encore faut-il être capable de la gérer. Dans le cas qui nous intéresse (l'analyse des génomes), les données telles que des séquences biologiques sont la matière première à partir de laquelle des programmes de calculs vont produire de nouvelles informations. Ces résultats sont à leur tour gérés pour être analysés par une expertise humaine. Enfin, cette expertise doit être validée de manière statistique et/ou expérimentale pour faire progresser la connaissance du monde vivant. Ce processus d'intégration et d'interrogation de données et de méthodes doit se faire à grande échelle car la quantité de séquences croît de manière exponentielle.

Bien que la confusion entre les notions de banque et de base soit aujourd'hui moins fréquente, il est important de rappeler ici ce qui les différencie. Les données contenues dans les banques sont distribuées sans outils d'organisation, et sous forme de fichiers de texte dans un format lisible par l'homme (fichiers à plat). Les bases de données intègrent trois éléments :

1. le système de gestion de base de données (le SGBD est le logiciel permettant l'exploitation de la base)
2. le modèle de données
3. les données distribuées, composées de fichiers texte dans un format lisible par la machine

Par exemple, dans le cas des SGBD relationnels, chaque ligne correspond à un *tuple* dont les champs sont délimités par un séparateur. Le SGBD permet d'instancier le modèle, c'est-à-dire de charger la base avec les données disponibles (voir p. 66).

Les biologistes ont à leur disposition une multitude de sources d'informations. On atteint actuellement [Morgat & Rechenmann, 2002] un nombre de banques et de bases de données biologiques de l'ordre du millier. Cependant, aucune source n'est exhaustive à ce jour, y compris la revue annuelle *Nucleic Acids Research*, dédiée entièrement aux banques et bases de données biologiques. Il est difficile de classer les banques et les bases. Selon les auteurs, le nom et le

nombre de catégories, ainsi que les banques et bases listées à l'intérieur de chacune d'elles, varient [Baxevanis, 2003, Discala *et al.*, 2000]. Nous nous limiterons ici aux banques et bases essentielles à notre travail.

2.1 Banques de données généralistes

Une banque généraliste est un dépôt (*repository*), où l'utilisateur peut soumettre une séquence et ses annotations. La communauté des microbiologistes dispose à l'heure actuelle de plus de 144 génomes procaryotes entièrement séquencés et 410 projets de séquençage sont en cours d'achèvement. La banque des génomes *online* GOLD ([Bernal *et al.*, 2001] Tab. 2.1 p. 62) permet de contrôler l'avancée des projets de séquençage dans le monde. Toutes ces séquences sont répertoriées dans des banques généralistes dont nous présenterons les plus utilisées.

2.1.1 Rappel historique

La séquence[1] est l'élément central[2] autour duquel la bioinformatique s'est développée. Les banques de données biologiques sont nées des avancées technologiques aussi bien dans le domaine de l'informatique que de la biologie. Une banque de données est un ensemble structuré de données non indépendantes. La première banque de données biologiques est une banque de séquences protéiques. En 1965, M. Dayhoff *et coll.* compile les premières séquences protéiques dans la banque *Atlas of Protein Sequence and Structure*[3] [Dayhoff *et al.*, 1965] utilisable par des outils informatiques de comparaison de séquences et permettant de comprendre l'évolution moléculaire. La seconde banque rassemble des structures macromoléculaires, et fut créée en 1971, c'est la *Protein Data Bank* [Bernstein *et al.*, 1977]. C'est encore aujourd'hui la principale banque de structures tridimensionnelles. Au début des années 1980, avec la découverte de la technique du séquençage des acides nucléiques[4], les premières grandes banques généralistes de séquences nucléiques voient le jour. C'est ainsi qu'apparaît la troisième banque biologique en 1979 : la *Los Alamos Sequence Library*, qui deviendra GenBank [Kanehisa *et al.*, 1984].

Les premières banques de séquences étaient de simples fichiers de données « à plat ». Chaque fichier contenait potentiellement plusieurs séquences, dont le nom servait d'accès aux programmes d'analyse. Le fichier était alors parcouru linéairement jusqu'à atteindre la séquence désignée (accès séquentiel). Aujourd'hui, la quantité de données est telle qu'il est préférable de coupler les banques à des systèmes d'interrogation/extraction dédiés aux séquences biologiques pour retrouver rapidement les données (accès direct par index, voir p. 66).

[1] En 1958, le chimiste anglais F. Sanger détermine la séquence du peptide insuline. La méthode couramment utilisée aujourd'hui pour le séquençage des peptides est la dégradation d'Edman (médecin suédois) publiée en 1967. La dégradation de Sanger est utilisée pour déterminer le N-terminal d'une séquence peptidique.

[2] Voir le dogme central de la biologie moléculaire.

[3] D'abord imprimé (jusqu'en 1978), l'Atlas est ensuite proposé sous forme électronique. Aujourd'hui, le centre *Protein Information Ressource* gère la *Protein Sequence Database* qui a succédé à l'Atlas et est accessible en ligne.

[4] En 1977, la méthode de Sanger et la méthode de Maxam et Gilbert permettent le séquençage de l'ADN.

2.1.2 Banques de séquences nucléiques

Les séquences et leurs annotations sont répertoriées dans différentes banques de séquences d'ADN. En 1999, les trois principales banques de séquences nucléiques, GenBank, EMBL-EBI et DDBJ, ont donné naissance à une collaboration internationale tripartite appelée *Internationnal Nucleotide Sequence Database* (INSD[5]) ([Benson *et al.*, 2004], [Kulikova *et al.*, 2004], [Miyazaki *et al.*, 2004] Tab. 2.1 p. 62). Les journaux scientifiques ne publient pas un article décrivant une séquence biologique si celle-ci n'a pas été préalablement déposée dans une de ces trois banques. La mise en forme des données est définie par l'ensemble des normes, standards et conventions nécessaires pour répondre aux besoins de représentation, d'échanges et de traitements des informations. L'entrée d'une séquence biologique est composée de lignes commençant par le code du type d'information contenu dans la ligne (ID = identification de l'entrée, AC = numéro d'accession ...). Chaque entrée commence par un identifiant unique, se termine par une ligne de fin '//' et se décompose en trois parties[6] :

1. L'en-tête contient un ensemble d'informations générales sur la séquence : un identifiant unique, un ou plusieurs numéros d'accession, son origine, sa description, des références bibliographiques et des commentaires.
2. Vient ensuite la liste des caractéristiques biologiques de la séquence et leurs qualificatifs[7]. Une caractéristique (*feature*) est décrite par un type (*feature key* : *CDS*, *gene*, *RBS* ...), une position sur le chromosome (*location*), et un ensemble de qualificatifs (*qualifier for feature key* : /note=,/gene=,/product= ...).
3. La dernière partie présente la séquence nucléique proprement dite (entête de la séquence et séquence).

Depuis 2000, le NCBI met en place un projet de collection de séquences de référence de génomes d'intérêt scientifique, *RefSeq* [Pruitt *et al.*, 2003]. Le but de cette collection est de fournir des jeux d'annotations complets, « nettoyés » et non redondants de séquences biologiques (ADN génomique, ARNm et protéines) telles que celles des génomes complets procaryotes[8]. En plus des éléments précités, un statut indique ici le niveau de *curation*. Le statut d'une séquence retrace les différentes étapes du processus de *curation*. On le repère dans le champ *COMMENT* de l'entrée : *genome annotation* (annotation automatique sans vérification), *predicted* (annotation de gènes de fonctions inconnues mais dont l'évidence est supportée par des projets de séquençage d'ADN complémentaires), *provisionnal* (annotation de gènes dont la plupart ont des fonctions connues ou inférées), *reviewed* (annotation provisoire révisée manuellement par un expert) et *completeness* (*curation* des annotations complètes achevée). Lors du processus de *curation*, certaines protéines prédites qui appartiennent à des familles d'orthologues conservées (*conserved hypothetical protein*), mais qui en

[5] `http://www.ncbi.nlm.nih.gov/projects/collab/`
[6] `ftp.ebi.ac.uk/pub/databases/embl/release/usrman.txt`
[7] `ftp.ebi.ac.uk/pub/databases/embl/doc/FTv6.doc`
[8] `ftp://ftp.ncbi.nih.gov/genomes/Bacteria`

DB	Category	Content	http	Reference
DNA Data Bank of Japan (DDBJ)	Major sequence repositories	All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration	http://www.ddbj.nig.ac.jp	Miyazaki-2004
EMBL			http://www.ebi.ac.uk/embl.html	Kulikova-2004
GenBank			http://www.ncbi.nlm.nih.gov/	Benson-2004
NCBI Reference Sequence Project		Non-redundant collection of naturally occurring biological molecules	http://www.ncbi.nlm.nih.gov/RefSeq/	Pruitt-2003
TPA*		A Third Party Annotation sequence is derived from primary sequence data currently found in the WWDDL	http://www.ncbi.nlm.nih.gov/Genbank/tpa.html	
Clusters of Orthologous Groups (COG)	Comparative Genomics	Phylogenetic classification of proteins from 43 complete genomes	http://www.ncbi.nlm.nih.gov/COG	Tatusov-2001
HOBACGEN*		Protein sequences of bacteria organized into families	http://pbil.univ-lyon1.fr/databases/hobacgen.html	Perriere-2000-a
MBGD		Microbial genome database for comparative genomic analysis	http://mbgd.genome.ad.jp/	Uchiyama-2003
IS database	Gene Expression	list of insertion sequences isolated from eubacteria and archae	http://www-is.biotoul.fr/is.html	
Stanford Microarray Database		Raw and normalized data from microarray experiments	http://genome-www.stanford.edu/microarray	Gollub-2003
Comprehensive Microbial Resource	Genomic Databases	Completed microbial genomes (CMR)	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	Peterson-2001
EcoGene		*E. coli* K-12 sequences	http://bmb.med.miami.edu/EcoGene/EcoWeb/	Rudd-2000
EMGlib		Completely-sequenced prokaryotic genomes	http://pbil.univ-lyon1.fr/emglib/emglib.html	Perriere-2000-b
GenoList		Completed microbial genomes	http://genolist.pasteur.fr/	Moszer-2002
GOLD		Information regarding complete and ongoing genome projects	http://igweb.integratedgenomics.com/GOLD/	Bernal-2001
HGT-DB		Putative horizontally-transferred genes in prokaryotic genomes	http://www.fut.es/~debb/HGT/	Garcia-Vallve-2003
Indigo*		Gene-related knowledge of *B. subtilis* and *E. coli*	http://195.221.65.10:1234/Indigo/	Nitschke-1998
Micado*		Completed microbial genomes and functional analysis of *B. subtilis*	http://topaze.jouy.inra.fr/cgi-bin/micado/index.cgi	Samson-2001
MIPS		Protein and genomic sequences	http://www.mips.biochem.mpg.de/	Mewes-2002
BIND	Interactions	Molecular interactions, complexes and pathways	http://bind.ca	Bader-2003
BioCyc	Metabolic Pathways and Cellular Regulation	Genome, metabolic pathways, transporters and gene regulation	http://biocyc.org/	Karp-2002
IntEnz		Enzyme nomenclature	http://www.ebi.ac.uk/intenz/	Fleischmann-2004
Kyoto Encyclopedia of Genes and Genomes (KEGG)		Metabolic and regulatory pathways	http://www.genome.ad.jp/kegg	Kanehisa-2002
LIGAND		Chemical compounds and reactions in biological pathways	http://www.genome.ad.jp/ligand/	Goto-2002
RegulonDB		Escherichia coli transcriptional regulation and operon organization	http://www.cifn.unam.mx/Computational_Genomics/regulondb/	Salgado-2001
GenPept*	Protein Databases	Fasta formatted translations extracted from WWDDL records	http://www.ncbi.nlm.nih.gov/Sitemap/index.html#Proteins	
GenProtEC		E. coli K-12 genome, gene products and homologs	http://genprotec.mbl.edu	Liang-2002
ooTFD		Transcription factors and gene expression	http://www.ifti.org/	Ghosh-1998
PIR-NREF		Non-redundant reference database	http://pir.georgetown.edu	Wu-2002
PIR-PSD		Comprehensive, annotated, non-redundant protein sequence databases	http://pir.georgetown.edu	Wu-2002
SWISS-PROT/TrEMBL		Curated protein sequences	http://www.expasy.ch/sprot	Boeckmann-2003
UniProt*		SWISS-PROT/TrEMBL + PIR International Protein Sequence Database Collaboration	http://www.expasy.uniprot.org/	Apweiler-2004
InterPro domains	Protein Motifs	Integrated documentation resource for protein families, domains, sites	http://www.ebi.ac.uk/interpro/	Mulder-2003
PDB	Structure	Structure data determined by X-ray crystallography and NMR	http://www.pdb.org/	Bourne-2004
Rfam	RNA Sequences	Non-coding RNA families	http://www.sanger.ac.uk/Software/Rfam/	Griffiths-Jones-2003
Ribosomal Database Project		rRNA sequence data, analysis tools, alignments, phylogenies (RDP-II)	http://rdp.cme.msu.edu	Cole-2003
tRNA sequences		tRNA and tRNA gene sequences	http://www.uni-bayreuth.de/departments/biochemie/trna/	Sprinzl-1998

TAB. 2.1 – Tableau des ressources procaryotes, modifié de A. Baxevanis [Baxevanis, 2003]

* Cette ressource ne figure pas dans le tableau original de A. Baxevanis.

FIG. 2.1 – Extraits de l'entrée GenBank du chromosome d'*E. coli* K-12(U00096.gbk)
La caractéristique, de type *CDS*, localisée en 362455 sur le brin inverse du chromosome d'*E. coli* K-12, contient l'information « lacZ » pour le qualificatif "/gene=" et « beta-D-galactosidase » pour le "/product=".

fait ne sont pas annotées dans GenBank, sont ajoutées à l'entrée. D'autres séquences protéiques sont modifiées, sur la base d'alignement avec des séquences d'autres espèces, pour corriger les décalages de phase et autres erreurs apparentes. Aujourd'hui, seuls deux génomes, *M. tuberculosis* H37Rv et *Y. pestis* CO92 ont le statut *completeness*.

Par ailleurs, la collaboration INSD a entrepris en 2002 la collection d'une nouvelle classe de données de séquences : les séquences consensus et l'annotation par un tiers *Third-Party Annotation and Consensus Sequences* (*TPA*[9]). Un enregistrement TPA est toujours lié à une publication montrant que les données (de séquences ou non) reposent sur un travail expérimental. Pour l'instant, un cinquantaine de CDS procaryotes sont présentes dans la nouvelle classe de données de séquence TPA[10].

Enfin, il existe d'autres banques de séquences nucléiques : par exemple, la banque des éléments génétiques mobiles (IS) impliqués dans le transfert horizontal chez les bactéries, la banque ribosomique (RDP) qui regroupe des données sur les séquences d'ARN ribosomique, ou encore la banque d'ARN de transfert qui compile les séquences des gènes d'ARNt de génomes complets ou publiés dans la littérature ([Cole *et al.*, 2003, Sprinzl *et al.*, 1998] Tab. 2.1 p. 62).

2.1.3 Banques de séquences protéiques

Après avoir obtenu un jeu de CDS de confiance pour un génome, on s'intéresse alors à la fonction des produits de leur traduction (les séquences protéiques correspondantes). Les recherches de similitudes reposent sur l'alignement de ces séquences protéiques avec celles des banques protéiques (Tab. 2.1 p. 62). Le format des fichiers à plat des banques protéiques ressemble à celui des banques nucléiques INSD. Il se divise en trois parties : l'en-tête, les *features* et la séquence, une entrée correspondant ici à un polypeptide. Les caractéristiques de la séquence protéique sont par exemple des modifications post-traductionnelles, des sites de liaison, des sites actifs, etc. On distingue plusieurs sources de séquences protéiques complètes et non redondantes (Tab. 2.1 p. 62).

TrEMBL-EBI correspond à la traduction de toutes les CDS de EMBL-EBI [Boeckmann *et al.*, 2003]. C'est une banque générée automatiquement, dans laquelle les séquences identiques sont regroupées. GenPept, maintenue par le NCBI, correspond à la traduction de toutes les CDS de GenBank. C'est une banque également générée automatiquement, mais les séquences identiques ne sont pas regroupées. La collaboration du PIR (*Protein Information Resource*), du MIPS (*Munich Information Center for Protein Sequence*) et du JIPID (*Japan International Information Database*), produit et distribue la banque *PIR-International Protein Sequence Database* (PIR-PSD) [Wu *et al.*, 2002]. Cette banque est le successeur de la banque *NBRF Protein Sequence* développée par M. O. Dayhoff. En 2001, PIR-PSD migre sous le SGBD Oracle©8i. Finalement, PIR-NREF est une banque non redondante de protéines de référence qui fournit une collection complète et à jour de toutes les données de séquences disponibles, notamment celles des projets de génomes complets [Wu *et al.*, 2002]. La banque contient toutes les données de PIR-PSD, Swiss-Prot,

[9] `ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb133.release.notes`
[10] `ftp://ftp.ebi.ac.uk/pub/databases/embl/tpa`

TrEMBL-EBI, *RefSeq*, GenPept et PDB [Bourne *et al.*, 2004], ainsi que les liens vers les différentes sources, tout en essayant de maintenir une redondance minimale.

Swiss-Prot, développée et maintenue au *Swiss Institute of Bioinformatics* (SIB), se distingue surtout par sa qualité d'annotation [Boeckmann *et al.*, 2003]. En effet, A. Bairoch, responsable de Swiss-Prot, a adopté une stratégie différente de celle de PIR-PSD : toute entrée de Swiss-Prot doit être vérifiée manuellement par un expert (rien ne rentre automatiquement dans SWISS-PROT). Cette restriction diminue le nombre d'entrées erronées ainsi que la redondance des séquences dans la banque, et augmente la qualité de l'annotation protéique. En revanche, Swiss-Prot contient deux fois moins d'entrées que PIR-PSD mais sa qualité la rend unique en son genre. Par ailleurs le SIB a récemment mis en place un projet d'annotation automatique et manuelle, de haute qualité, des protéomes microbiens : le projet *HAMAP* ([Gattiker *et al.*, 2003] voir p. 152). Au niveau européen, il existe une collaboration entre Swiss-Prot et l'EBI. La banque de séquences protéiques complète et non redondante SWALL (SP_TR_NRDB) combine deux banques complémentaires : Swiss-Prot et TrEMBL (Swiss-ProtTrEMBLet TrEMBLnew) et a donc l'avantage d'être exhaustive.

Enfin, sur le plan international, Swiss-Prot, TrEMBL-EBI et PIR-PSD ont été unifiées pour former la base universelle de connaissance protéique (*Universal Protein Knowledgebase*, *UniProt* [Apweiler *et al.*, 2004]). Ceci afin de fournir à la communauté scientifique une ressource de référence pour l'annotation fonctionnelle des séquences protéiques, unique et centralisée.

2.1.4 Problèmes posés par les banques

Le problème central posé par les banques de séquences est lié au *parsing* (ou analyse syntaxique) des annotations stockées dans les fichiers à plat. On cherche à découper les annotations en informations élémentaires pour les utiliser à bon escient, par exemple pour les structurer dans une base de données, comme nous le verrons par la suite. La difficulté, pour récupérer ces informations, est liée en grande partie à l'hétérogénéité des annotations. Bien qu'il existe une norme, *The DDBJ - EMBL-EBI - GenBank Feature Table Definition*[11] (définition INSD de la liste des caractéristiques d'une séquence, version 6.0), chaque annotateur l'interprète et l'utilise à sa manière. Il persiste donc une hétérogénéité dans les annotations aussi bien sur le fond (les annotations de certains chromosomes sont plus riches et plus avérées que celles d'autres chromosomes), que sur la forme (une même information *valeur* peut être rangée à différents endroits *attribut*). On comprend qu'il est donc difficile pour le bioinformaticien d'extraire correctement les données de ce genre de fichier. Pour ne citer que quelques exemples :

- Pour renseigner le qualificatif /product, les critères de similitudes et les seuils choisis restent obscurs et sont différents selon les annotateurs : *unknown*, *doubtful*, *putative*, *possible*, *probable*, *very hypothetical*, *hypothetical*, *conserved hypothetical*, etc.
- Le qualificatif /gene peut contenir un nom de gène (par exemple *lacZ*), éventuellement suivi de noms synonymes ou bien l'étiquette du gène (b0343) qui est en fait une étiquette unique à

[11] `http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html`

l'objet CDS. Pour lever cette confusion entre étiquette et nom de gène, la collaboration INSD a décidé de créer un nouvel attribut[12] : /locus_tag.

– Le qualificatif /note a tendance à devenir un attribut « fourre-tout ». Selon les annotateurs, il ne contient pas la même information, et est de ce fait difficile à analyser automatiquement. On peut y trouver des résultats sur la meilleure caractéristique extrinsèque (*Similar to ...*) et/ou sur des caractéristiques intrinsèques de la séquence (*low GC*, *this region contains an authentic frameshift ...*) et/ou des informations sur les noms synonymes de gènes (*alternate gene name : ...*).
– Un autre problème important réside dans la manière d'annoter les pseudogènes et CDS contenant des mutations et erreurs de séquence entraînant des décalages de phase ou du cadre de lecture (*frameshift*). Plusieurs illustrations de cette constatation sont données dans la figure 2.2 p. 67. Notamment cette hétérogénéité d'annotation des décalages du cadre de lecture est à l'origine d'une différence importante entre les fichiers GenBank et EMBL-EBI. Généralement, dans les fichiers GenBank chaque objet CDS est associé à un objet gene alors qu'il n'y a pas d'objet gene dans les fichiers EMBL-EBI. Or, il arrive que dans certains fichiers GenBank, les pseudogènes soient annotés uniquement par un objet gene (pas d'objet CDS associé), dans ce cas, ces pseudogènes seront absents du fichier EMBL-EBI correspondant. Par exemple, dans le fichier NC_004741.gbk au format GenBank des annotations du chromosome de *Shigella flexeneri* 2a 2457T, il existe 295 pseudogènes (objet gene possédant l'attribut /pseudo) qui sont absents du fichier correspondant au format EMBL-EBI, AE014073.embl.

Finalement, derrière les problèmes d'hétérogénéité d'annotation, nous découvrons des problèmes d'ontologie. Tout le monde ne met pas la même définition derrière un mot et tous les mots que l'on voudrait utiliser ne sont pas forcément définis par le standard INSD. Par exemple, dans le format GenBank, tout objet *CDS* est associé à un objet *gene* (ce qui n'est pas le cas du format EMBL-EBI). Or, selon le standard INSD, le type *gene* définit l'unité de transcription, ce qui est erroné dans le cadre d'un opéron polycistronique. On peut se demander si la création d'un nouveau type *operon* ne serait pas préférable. Ainsi chez les procaryotes, le type *gene* définirait alors l'unité de traduction (l'ORF). Bien que les banques de séquences soient incontournables pour rassembler les données et les mettre à la disposition de la communauté scientifique, l'utilisation efficace des données d'annotation reste limitée. C'est pourquoi on observe une migration d'une partie de ces données (génome complet, thématique biologique, etc.) vers des bases de données.

2.2 Bases de données et leurs systèmes de gestion

2.2.1 Rappel historique

Les années 60 connaissent un premier développement des bases de données sous forme de systèmes de gestion de fichier (Fig. 2.3 A p. 68). La première génération de Système de Gestion de

[12] `ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb133.release.notes`

FIG. 2.2 – Décalages du cadre de lecture choisis dans les entrées GenBank des chromosomes complets de *Y. pestis* CO92 et de *Pseudomonas aeruginosa*

A) Le décalage du cadre de lecture est indiqué par l'utilisation de *join* et /pseudo. La CDS n'est pas traduite (/translate n'est pas renseigné).

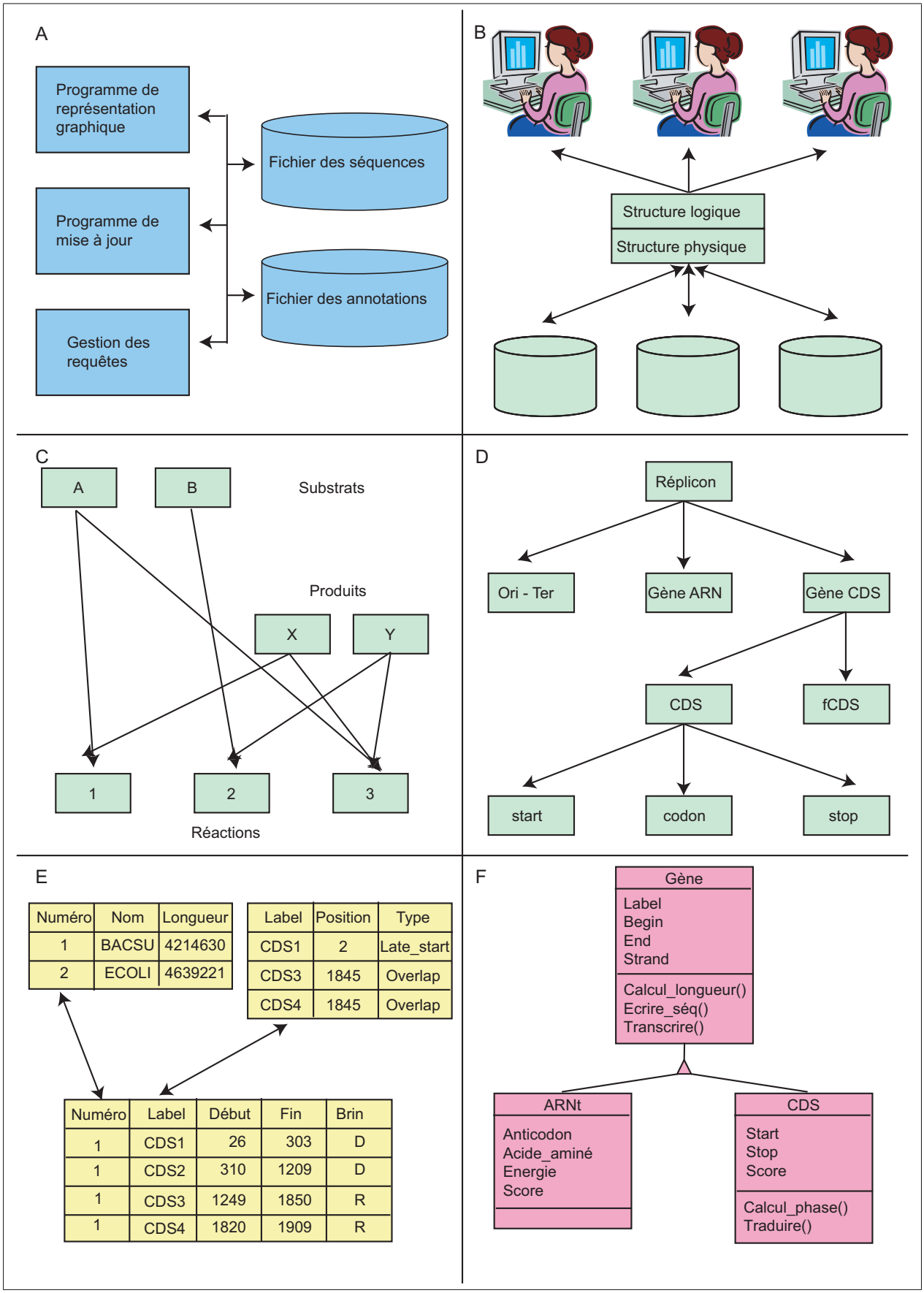
B) Le décalage du cadre de lecture est indiqué par l'utilisation de /pseudo. La CDS n'est pas traduite.

C) Le décalage du cadre de lecture est indiqué par l'utilisation de *join*. La CDS est traduite (/translate est renseigné).

D) Le décalage du cadre de lecture est indiqué par l'utilisation de /partial. La CDS est traduite.

E) Le décalage du cadre de lecture n'est pas indiqué par *join*, /pseudo ou /partial. Dans le cas de l'annotation GenBank, il n'y a pas d'objet CDS donc pas de traduction. Dans le cas de l'annotation *RefSeq*, les deux CDS sont traduites.

F) Le décalage du cadre de lecture n'est pas indiqué par *join*, /pseudo ou /partial. Il n'y a pas d'objet CDS donc pas de traduction.



Numéro	Nom	Longueur
1	BACSU	4214630
2	ECOLI	4639221

Label	Position	Type
CDS1	2	Late_start
CDS3	1845	Overlap
CDS4	1845	Overlap

Numéro	Label	Début	Fin	Brin
1	CDS1	26	303	D
1	CDS2	310	1209	D
1	CDS3	1249	1850	R
1	CDS4	1820	1909	R

FIG. 2.3 – SGBD

FIG. 2.3 – SGBD

A) SGBD précurseur En 1965, les précurseurs des systèmes de gestion de base de données (SGBD) modernes sont des systèmes de gestion de fichiers à plat (*file system*). Les fichiers sont reliés par des références logiques mises en œuvre par des pointeurs. Ils sont composés d'articles, stockés les uns à la suite des autres et accessibles par des valeurs de données appelées clés (identifiant unique de l'article ou clé primaire). Ainsi, on accède à un article par sa clé en parcourant des chaînes d'articles entre fichiers (accès séquentiel). Il n'y a pas de modèle de données. La structure des données est décrite dans les programmes de traitement. Chaque modification de la structure des données entraîne la modification de tous les programmes.

B) SGBD première génération ou SGBD légataire En 1970, la première génération de SGBD est marquée par la séparation de la description des données et de la manipulation par les programmes d'applications. Tout SGBD doit satisfaire aux requis suivants : persistance des données, gestion du disque, fiabilité et reprise après panne, partage des données et interrogation ad hoc multi-utilisateur. Pour garantir un accès rapide à l'information, un fichier d'index permet d'associer une clé d'un article à son adresse relative dans le fichier à plat. Ces SGBD de première génération sont aujourd'hui appelés systèmes légataires car ils nous ont été légués par le passé. Le modèle de données est directement issu de la modélisation d'organisation d'articles stockés dans des fichiers et reliés entre eux par des pointeurs. Ces modèles sont aujourd'hui inadaptés en tant que modèle conceptuels de données. Ils assurent en effet une faible indépendance des programmes aux données.

C) SGBD réseau C'est le modèle de données réseau des SGBD première génération. Les nœuds du graphe représentent des types de données. Les arcs du graphe représentent des types de pointeurs reliant les différents types de données.

D) SGBD hiérarchique C'est le modèle de données hiérarchique des SGBD première génération. Le modèle hiérarchique peut être vu comme un cas particulier du modèle réseau, l'ensemble des liens entre types d'articles devant former des graphes hiérarchiques. Les bases de données modélisent des informations du monde réel. Puisque le monde réel nous apparaît souvent au travers de hiérarchies, il est normal qu'un des modèles les plus répandus soit le modèle hiérarchique.

E) SGBD deuxième génération ou SGBD relationnel Dans un SGBD légataire, l'information est physiquement répartie en plusieurs lieux de la mémoire, reliés par le biais de pointeurs. En 1980, les SGBD deuxième génération ou SGBD relationnels (SGBDR) généralisent la décomposition de l'information concernant un même objet, en imposant la factorisation des informations communes selon leur sémantique. Par exemple, les informations concernant les chromosomes sont stockées ensemble ; les informations concernant les décalage du cadre de lecture aussi. Les CDS qui constituent une relation liant les chromosomes aux décalages du cadre de lecture apparaissent en un troisième lieu (une troisième table). Cette organisation sémantique et hiérarchique s'appuie sur la théorie mathématique des relations. Ainsi, au moins trois objectifs sont atteints par les SGBDR :

- Premièrement, la clarification des problèmes de cohérence et de redondance. En effet, la factorisation réduit la répétition et la théorie logique des relations permet de définir les règles d'intégrité (clé primaire unique, valeur nulle, clé étrangère) qui décrivent les contraintes que doivent satisfaire les données correctement saisies.
- Deuxièmement, l'indépendance des implémentations de la structure des données. En effet, la représentation des données, sous forme de table de valeurs communes ou relation, est simplifiée et permet donc une meilleure interface.
- Troisièmement, l'implémentation de logiciels de manipulation logique des données (et non plus procédurale) à partir du support théorique de l'algèbre relationnelle (l'union, la différence, le produit cartésien, la restriction, la projection et la jointure), on définit le standard *Structured Query Langage* (*SQL*) puis on réalise des implémentations (MySQL, PostgreSQL, Oracle, Sybase). Le modèle relationnel qui peut être vu comme l'interprétation d'un langage logique devient donc le standard des SGBD.

F) SGBD troisième génération ou SGBD objet En 1990, les SGBD de troisième génération ou SGBD objet (SGBDO) regroupent deux mondes : celui de la programmation orientée objets (langages comme Smalltalk, C++ et Java) et celui de la représentation des connaissances. Ils regroupent les concepts essentiels pour modéliser de manière progressive des objets complexes encapsulés, intégrant les opérations de manipulation qui les définissent pragmatiquement. Un *objet* est l'abstraction informatique d'une entité du monde réel. Il lui correspond une identité, ou identifiant, un état (décrit par un ensemble d'attributs) et un ensemble de comportements (les opérations applicables à l'objet et modifiant éventuellement son état). Une classe est l'abstraction décrivant un type d'objets. Tout objet est donc une instance d'une classe. Une classe est définie par la liste de ses attributs et celle de ses méthodes. Une relation d'*héritage* peut être définie entre certaines classes. Cette relation permet la mise en place automatique de la transmission d'une partie des propriétés d'une classe vers une sous-classe. Les propriétés des classes sont l'héritage, le polymorphisme l'encapsulation. L'*héritage* permet de spécialiser une classe générale (comme gène) en une classe plus précise (comme ARNt ou CDS). L'*encapsulation* groupe et occulte l'ensemble des données et des procédures associées à un objet, ne laissant visible qu'une interface composée des opérations effectivement accessibles aux entités pouvant communiquer avec lui. Le *polymorphisme* permet à une même opération de s'appliquer à des objets de différentes classes. Ainsi *trier*() est une opération à vocation polymorphe puisqu'on voudra l'utiliser pour des objets de la classe *gene* comme pour des objets de la classe *carte*. Il permet aussi d'appliquer à des objets d'une même classe une même opération dont les paramètres peuvent être de types différents. Comme par exemple *trier*() par ordre croissant, ou bien décroissant. Enfin certains attributs peuvent être multivalués, par exemple un livre peut avoir plusieurs auteurs. Regrouper plusieurs objets pour former un seul attribut s'effectue au moyen de classes génériques permettant de supporter des *collections* d'objets. Comme pour les SGBDR, les SGBDO sont implémentés à l'aide de langages standard. L'*Object Definition Language* (*ODL*) est le langage de définition de schéma de base de données objet. L'*Object Manipulation Language* (*OML*) est le langage de manipulation des objets (gestion de la persistance des objets). L'*Object Query Language* (*ODL*) est le langage d'interrogation des bases de données objet. Il étend l'algèbre relationnelle aux objets.

Base de Données (SGBD ; Fig. 2.3 B, C et D p. 68) permet en plus de modéliser les données et d'accéder plus rapidement à l'information contenue dans les fichiers à plats au moyen d'index. Ainsi, aux fichiers à plat des banques de données sont associés des systèmes d'interrogation dédiés comme *Entrez*, ou *SRS* [Frishman *et al.*, 1998a], ou encore *ACNUC*[13]. La base *ACNUC*, développée par M. Gouy *et coll.* (1985) est un SGBD qui sépare la structure logique des données (modèle entité/association) de la structure physique. La principale limitation d'ACNUC est liée au fait que ce système ne permet d'interroger qu'une seule banque à la fois (*e.g.* EMBL-EBI, GenBank, SWALL, PIR-PSD). De plus, il existe une très forte dépendance entre les données et le langage d'interrogation. De ce fait, toute modification dans la façon dont sont organisées les données va obliger le concepteur à modifier le système d'interrogation. En contrepartie, du fait même de cette dépendance, les requêtes sont généralement plus puissantes qu'avec des systèmes de bases de données de deuxième génération. Avec les SGBD de deuxième génération (SGBD relationnels, SGBDR ; Fig. 2.3 E p. 68) puis de troisième génération (SGBD objets, SGBDO ; Fig. 2.3 F p. 68), les données deviennent indissociables du SGBD [Durand *et al.*, 2003]. En effet, on n'accède plus aux données au travers de fichiers à plat mais directement dans les tables de la base. En 1984, la banque de séquences nucléiques Los Alamos est la première à migrer vers un SGBDR. La distribution des données est toujours réalisée sous forme de fichiers à plat.

Le concept objet a suscité un vif intérêt dans la communauté des bases de données, comme pour son analogue dans la programmation. Cependant, le succès des SGBDO est aujourd'hui encore limité par les problèmes liés à la persistance, l'optimisation des requêtes, la modification du schéma, etc.). Ainsi les SGBD relationnel-objet (SGBDRO) ont vu le jour dans les années 2000. Par exemple, iProClass (successeur de ProClass ; Tab. 2.1 p. 62) est une base de séquences, de familles, de structures et de fonctions protéiques implémentée aux Etats-Unis en 2001 sous le SGBDRO Oracle 8i.

2.2.2 Exemples de bases de données spécialisées

L'hétérogénéité et la redondance des séquences des banques généralistes ont conduit au développement de bases spécialisées autour de thématiques particulières. Le nombre de bases de données spécialisées disponibles aujourd'hui est très important [Baxevanis, 2003, Discala *et al.*, 2000]. Afin de simplifier cette présentation nous les avons rassemblées en deux catégories principales : les bases de données thématiques et les bases de données génomiques.

Les *bases de données thématiques* rassemblent les données relatives à une thématique biologique particulière : compilation de motifs protéiques structuraux, définition de familles de transporteurs, analyse de réseaux de régulation et de réseaux métaboliques, etc.

– Il existe au moins une quinzaine de banques et bases de motifs de séquences protéiques. Dans un souci d'homogénéisation, la base Interpro regroupe les entrées de PROSITE [14], PRINTS[15],

[13] `http://pbil.univ-lyon1.fr/databases/acnuc.html`

[14] Familles de *pattern* pour les domaines protéiques et sites fonctionnels ; familles de profils.

[15] Familles de *fingerprints* protéiques : groupe de motifs plus performant dans un contexte biologique qu'une repré-

Pfam[16], Prodom[17], SMART[18], TIGRFAM[19], et superfamilles PIR-NREF ([Mulder *et al.*, 2003] Tab. 2.1 p. 62). L'ensemble des entrées de cette banque de domaines protéiques, au nombre de 8423 (version 6.2) ont été vérifiées manuellement (signification biologique, référence bibliographique). C'est une ressource indispensable pour l'analyse de protéomes, particulièrement pour l'annotation fonctionnelle de protéines multidomaines.

– La base IntEnz (*Integrated relational Enzyme database* [Fleischmann *et al.*, 2004]) développée par l'EBI, regroupe les données enzymatique de trois sources différentes :

  1. la liste des enzymes approuvés par le Comité de Nomenclature de l'Union Internationale de Biochimie et Biologie Moléculaire (NC-IUBMB [IUBMB, 1992])
  2. la banque de nomenclature des enzymes du SIB (ENZYME [Bairoch, 2000])
  3. la base de fonctions enzymatiques de l'université de Cologne (BRENDA [Schomburg *et al.*, 2004]).

  IntEnz contient, pour chaque enzyme, le numéro EC (*Enzyme Commission*), son nom, ses cofacteurs, la réaction catalysée, etc.. Partant de ces données biochimiques, plusieurs bases thématiques décrivant les voies métaboliques d'un ou plusieurs organismes ont été développées, notamment les bases BioCyc de P. Karp (EcoCyc pour ECOLI et sa généralisation multigénomes MetaCyc [Krieger *et al.*, 2004]), ou encore la base de l'université de Kyoto KEGG [Kanehisa *et al.*, 2004]. Dans le cas d'*E. coli* K-12, il existe aussi une base contenant les données de régulation transcriptionnelle, et d'organisation en opéron RegulonDB ([Salgado *et al.*, 2004] Tab. 2.1 p. 62).

– L'ultime étape de l'annotation fonctionnelle est la validation expérimentale. Elle permet de démontrer l'existence du produit d'un gène. On peut trouver cette information dans des bases de données d'expression résultant de la technique des puces à ADN (*microarray*, *DNA chip*), bases trop peu nombreuses à ce jour. Seule la base de Stanford, *Stanford Microarray Database*, fournit des données d'expression sur les bactéries *B. subtilis*, *Campylobacter jejuni*, *E. coli* K-12, *H. pylori* 26695, *Salmonella enterica* serovar Typhimurium LT2 et *Streptomyces coelicolor* ([Gollub *et al.*, 2003] Tab. 2.1 p. 62).

Les *bases génomiques* permettent de réunir les séquences d'*une même espèce* et d'en mettre à jour les annotations et sont, à ce jour, très nombreuses :

– En 1990, une des premières bases spécialisées, *Colibri*, est dévéloppée pour compiler les séquences de la bactérie *E. coli* K-12 [Médigue *et al.*, 1990]. ACeDB (A *Caenorhabditis elegans* DataBase ; Tab. 4.1 A p. 154) est un SGBD orienté objet dédié aux projets de séquençage, conçu aussi en 1990 par l'EBI. Partant de la structure de base de données relationnelle de la base Colibri, un modèle générique appelé GenoList a été défini [Boneca *et al.*, 2003].

sentation par un seul motif.

[16] Profils HMM.

[17] Familles de domaines protéiques.

[18] Profils HMM.

[19] Profils HMM.

Sont construites sur ce modèle Les bases génomiques de l'Institut Pasteur de Paris : *Colibri*, *Subtilist*, *TubercuList*, etc. Par ailleurs, pour *E. coli* K-12, il existe aussi la base EcoGene [Rudd, 2000] qui fournit un jeu vérifié d'annotations syntaxiques et fonctionnelles, et la base *GenProtEC*, qui contient une mise à jour très régulière de l'annotation fonctionnelle du protéome de cette bactérie ([Serres *et al.*, 2004] TAB. 2.1 p. 62).

- D'autres bases spécialisées tendent à rassembler les annotations de plusieurs génomes au sein d'une même structure : c'est le cas de la base *EMGlib* (*Enhanced Microbial Genomes Library* [Perrière *et al.*, 2000a]) qui rassemblent les génomes procaryotes provenant de INSD. Micado, qui repose sur la base EMGlib, est centrée sur l'analyse fonctionnelle de *B. subtilis* ([Biaudet *et al.*, 1997] TAB. 2.1 p. 62).
- La base multigénomes du TIGR, CMR (*Comprehensive Microbial Resource*, [Peterson *et al.*, 2001] TAB. 2.1 p. 62) contient les données d'annotation de INSD et une réannotation automatique générée par les outils d'analyse utilisés au TIGR[20].
- Plusieurs autres bases spécialisées multigénomes incluent des relations d'homologie, d'orthologie, de paralogie, de synténie. La banque de donnée de COG (*Clusters of Orthologous Groups of proteins* [Tatusov *et al.*, 2001]), développée au NCBI, réalise une classification phylogénétique et fonctionnelle des protéines de génomes microbiens complets, et plus récemment de quelques génomes eucaryotes. Un COG est un groupe de gènes supposés orthologues, un même gène pouvant appartenir à un ou plusieurs COG. La base MBGD (*Microbial Genome Database* [Uchiyama, 2003]) permet d'identifier des groupes d'orthologues, de paralogues, et de comparer l'ordre des gènes entre plusieurs génomes. La table des groupes de gènes orthologues est créée dynamiquement à partir des résultats de similitudes précalculés et stockés dans la base.
- Enfin HOBACGEN [Perrière *et al.*, 2000b] est une base contenant des familles de protéines similaires pour les protéomes procaryotes. Elle permet de sélectionner des jeux de gènes homologues et de visualiser les alignements multiples et les arbres phylogénétiques. Elle est donc utile pour la génomique comparative et l'évolution moléculaire chez les procaryotes. Elle est basée sur les données de la SWALL.

2.2.3 Limites des bases de données actuelles

Depuis l'émergence des banques (fichier de texte) et des bases (SGBDR, SGBDRO), chaque groupe de travail a pu laisser libre cours à son imagination pour mettre en forme, à sa manière, les données de ses axes de recherche. De ce fait, ces données sont très hétérogènes et surtout distribuées. Aussi des collaborations sont mises en place pour uniformiser les formats et regrouper les données afin de les mettre à la disposition de la communauté scientifique. Ce travail implique une collaboration étroite entre plusieurs petites équipes travaillant sur des bases complémentaires, et le développement d'un seul système regroupant toutes les données. Nous avons vu précédemment

[20] `http://www.tigr.org/software/`

le regroupement de banques et bases de domaines protéiques, réalisé dans InterPro sous Oracle. D'autres projets sont en train de voir le jour comme le SGBDRO Oracle 8i et UniProt[21] (Tab. 2.1 p. 62).

Toute base de données doit être accompagnée d'outils nécessaires pour la construire, l'interroger, la maintenir et la gérer, mais au cours de son utilisation des besoins nouveaux apparaissent en même temps que de nouvelles données. On voit ainsi très vite la nécessité d'intégrer de nouveaux outils d'analyse, et de produire, à partir de la source initiale de données, de nouvelles connaissances. Aussi la recherche dans ce domaine s'est elle orientée vers des modèles de bases de connaissances, ces bases contenant cette fois-ci des mécanismes d'inférence de nouvelles connaissances (*SHIRKA* [Rechenmann & Uvietta, 1991]). Aujourd'hui, plusieurs plates-formes d'analyse ou d'annotation des génomes (encore appelées *workbench system*) sont développées et intègrent à la fois la notion de bases de données (ou bases de connaissances) et des stratégies d'analyse très diverses. Nous reviendrons sur ces plates-formes au cours du chapitre 4.

[21] La collaboration internationale PIR/EBI/Swiss-Prot est née récemment, et le projet des bases protéiques unies, *United Protein Databases*, est en cours d'élaboration. Ce projet permettra de créer une banque centrale de séquences protéiques et de fonctions en groupant les forces de Swiss-Prot, TrEMBL et PIR.

Chapitre 3

Prédiction de gènes dans les séquences procaryotes et analyse de leur usage des codons

Le présent travail de recherche a pour objectif final l'annotation relationnelle d'îlots de pathogénie bactériens, ce qui implique nécessairement de disposer de jeux d'annotations « curés » (génomes et protéomes). Nous avons vu que les annotations des banques étaient hétérogènes (voir p. 65), ce qui nous a conduits à mettre en place des processus d'annotation et de réannotation des génomes procaryotes qui seront développés dans la seconde partie du présent mémoire.

Du point de vue des objectifs biologiques, on peut distinguer trois niveaux d'annotation des génomes procaryotes, de complexité croissante :

1. l'annotation syntaxique concerne l'identification d'objets génomiques sur le chromosome (*features*)
2. l'annotation fonctionnelle consiste à attribuer une (ou plusieurs) fonction(s) biologique(s) à ces objets et/ou à leur(s) produit(s) (*qualifiers*)
3. l'annotation relationnelle permet de définir des relations entre ces objets et/ou leur(s) produit(s) (voir p. 163)

L'annotation syntaxique est au centre de ce travail car elle représente les fondations du processus d'annotation. Certaines erreurs d'annotation syntaxique, comme par exemple un « faux » gène, vont se propager dans les annotations fonctionnelle et relationnelle. De ce fait, on constate que, d'une façon générale, les génomes procaryotes sont « sur-annotés » [Ochman, 2002, Lawrence, 2003, Skovgaard *et al.*, 2001].

Après une brève introduction sur les méthodes bioinformatiques, les deux sections suivantes décrivent des méthodes d'annotations syntaxiques des génomes procaryotes, plus particulièrement des méthodes de prédiction de gènes. La dernière section de ce chapitre est consacrée aux méthodes statistiques d'analyse de données qui nous ont permis, ici, d'étudier l'usage du code génétique de

génomes bactériens (annotation fonctionnelle). Comme nous le verrons, un des objectifs de cette étude est d'améliorer les programmes de prédiction de gènes en tenant compte de l'hétérogénéité en oligonucléotides des gènes d'un chromosome. La plupart des méthodes décrites ici se fondent sur l'existence de biais compositionnels dans les séquences biologiques, notion que nous avons introduite dans le premier chapitre.

3.1 Introduction aux méthodes bio-informatiques

Dès 1965 est apparue la discipline des biomathématiques ou biostatistiques pour répondre aux besoin de la phylogénie moléculaire. Il n'est donc pas étonnant que la première méthode d'analyse informatique de séquences soit une méthode de construction d'arbre phylogénétique [Fitch, 1970]. Puis sont apparues les méthodes de recherche de similitudes par programmation dynamique avec évaluation de la qualité de l'alignement, d'abord l'alignement global de deux séquences [Needleman & Wunsch, 1970], puis l'alignement local [Smith & Waterman, 1981]. Ensuite, ce fut le tour des méthodes multifactorielles pour décrire l'usage du code génétique (*codon usage*) [Grantham *et al.*, 1981]. Enfin, depuis plus de 10 ans, les bioinformaticiens se tournent vers des modèles probabilistes plus complexes (modèles markoviens) pour décrire les phénomènes biologiques [Borodovsky & McIninch, 1993b, Krogh *et al.*, 1994]. Cette reformulation (*model-driven approach*) sert à résoudre des problèmes fondamentaux de la bioinformatique : alignement de séquences, prédiction de structure, phylogénie moléculaire, détection de gènes, etc.

Aujourd'hui, il devient difficile d'établir une classification des différentes méthodes d'analyse de séquences pour plusieurs raisons. D'abord, il existe un nombre incalculable de méthodes issues de domaines très variés (chimie, physique, statistique, intelligence artificielle, etc) [Koonin, 2002, Mount, 2001]. Ces méthodes furent un temps répertoriées dans le Biocatalog[1] qui n'est plus maintenu depuis juillet 2000. Ce n'est qu'en juillet 2003 que la revue *Nucleic Acids Research* consacre, pour la première fois, un numéro entier aux serveurs *web* dédiés à l'analyse de séquences et l'étude de structures biologiques, alors que ce même journal sort un numéro spécial « Bases de données biologiques » chaque année depuis plus de dix ans. Néanmoins, il existe des serveurs publics de logiciels bioinformatiques dont les mises à jour, le contenu et l'ergonomie sont de qualité variable. Parmi les plus connus on peut citer :

Services sur le Web	Infobiogen	`http://www.infobiogen.fr/services/menuserv.html`
Logiciels pour la biologie	Institut Pasteur	`http://bioweb.pasteur.fr/`
Archives pour la biologie	IUBio	`http://iubio.bio.indiana.edu/`
Logiciel de bioinformatique	Université d'Oxford	`http://www.molbiol.ox.ac.uk/`
Serveur de protéomique	SIB	`http://us.expasy.org/`
Logiciels « Open Source »	SourceForge	`http://sourceforge.net/softwaremap/trove_list.php?form_cat=252`

[1] Le *biocatalog* (`http://www.ebi.ac.uk/biocat/biocat.html`) est une banque de logiciels spécialisés dans la biologie moléculaire et génétique.

Le choix d'une méthode d'analyse de séquence est conditionné par un critère essentiel : l'évaluation en fin de calcul de la significativité[2] des résultats (un score). Il existe principalement deux types d'approche pour évaluer la significativité d'une prédiction : le test statistique (*log-likelihood ratio*, P-value, E-value) et le théorème de Bayes.

Un test statistique simple est celui du *log-likelihood ratio* ou *log-odds ratio* [Durbin *et al.*, 2001e, Altschul, 1991], pour lequel on compare les probabilités de séquences sous deux modèles différents, comme le modèle de séquences biologiques et le modèle de séquences aléatoires (modèle $M0$ d'indépendance entre les positions). Cette approche est utilisée dans le cadre des modèles de séquences codantes ou non-codantes [Guigo, 1999] et des modèles de paires de séquences similaires ou non similaires (score d'alignement ; [Durbin *et al.*, 2001e]). Un autre test consiste à utiliser la distribution des valeurs extrêmes [Durbin *et al.*, 2001e]. Par exemple, lors d'une recherche de similitude entre une séquence requête et des séquences biologiques d'une banque, après avoir déterminé les alignements et calculé leur score (S_a), on évalue la probabilité d'avoir au moins un alignement de score supérieur ou égal à S_a dans une banque de séquences aléatoires (P-value ; [Bejerano, 2003]). Ainsi, quelque soit le type d'alignement obtenu (global ou local), déterminer sa significativité statistique consiste à se demander si le score obtenu est suffisamment élevé pour ne pas être imputé au hasard. Pour chaque alignement, le programme Blast calcule la E-value, c'est-à-dire le nombre espéré d'alignements de score supérieur ou égal à S_a dans une banque de séquences aléatoires. Cette mesure est préférée à la P-value car il est plus facile de comprendre la différence entre une E-value de 5 et 10, qu'entre une P-value de 0,993 et de 0,99995. Cependant, quand la E-value est inférieure à 0,01, P-value et E-value sont quasiment identiques.

On peut reprocher au rapport de log-vraisemblance de ne pas vraiment représenter ce que l'on cherche à évaluer. En effet, il permet de comparer la probabilité d'une séquence sous deux modèles différents ($\mathbb{P}(X \mid M)$), or on cherche à prédire le modèle qui s'ajuste le mieux sur la séquence, ce qui reviens à comparer les probabilités des modèles connaissant la séquence ($\mathbb{P}(M \mid X)$) [Durbin *et al.*, 2001e]. L'approche Bayesienne permet d'inverser une probabilité conditionnelle, en calculant la probabilité *a posteriori* d'un événement en fonction des probabilités *a priori* de cet événement : $\mathbb{P}(M \mid X) = (\mathbb{P}(X \mid M) * \mathbb{P}(M))/(\mathbb{P}(X))$ [Durbin *et al.*, 2001b].

3.2 Prédiction de gènes spécifiant des ARN fonctionnels

Une fois déterminée, la séquence génomique ne représente que la donnée brute : il faut la déchiffrer en utilisant une grammaire dans laquelle opérateur, promoteur, ARNt, ARNr, RBS, CDS, terminateur, etc, se combinent d'une manière non définitive selon des règles établies à partir des connaissances en biologie moléculaire et d'outils bioinformatiques : cette combinaison permet de constituer l'unité fonctionnelle qu'est le gène. Selon l'origine de la séquence génomique, procaryote ou eucaryote, la méthodologie employée sera différente. Les problèmes posés par l'identification des gènes sont distincts dans le cas des génomes très peu denses en régions codantes (chez les

[2] Un événement significatif a vraiment peu de chance d'être dû au hasard, par définition, il est exceptionnel.

mammifères) ou dans le cas des génomes microbiens (10% de non-codants contre 90% de séquences transcrites). Ces différences entraînent l'utilisation d'algorithmes distincts, dont la conception évolue à la lumière des progrès de la biologie. Par exemple, la découverte des mécanismes d'épissage permet de mieux prédire les sites d'épissage : on ne recherche plus simplement GU pour prédire le site donneur, mais aussi son contexte en $3'$ dont le consensus est AAGT. Inversement, les prédictions bioinformatiques devraient aussi permettre d'influencer les expérimentations biologiques [Grantham *et al.*, 1980, Ikemura, 1985].

3.2.1 ARN de transfert

L'annotation syntaxique d'un ARNt consiste à définir ses bornes. Les gènes d'ARNt ont tendance à avoir une structure secondaire (bases appariées) plus conservée que leur séquence primaire [Lowe & Eddy, 1997]. C'est pourquoi les méthodes de prédiction d'ARNt sont fondées sur les propriétés intrinsèques des séquences afin de déterminer si un repliement en ARNt est probable.

Une première approche utilise la recherche de signaux de séquence linéaire à l'aide de matrices de type *log-odds* décrivant des motifs spécifiques des ARNt (boîtes promotrices A et B internes aux bras TPsiC et D, distance entre ces boîtes, distance au signal poly-T de terminaison de transcription). L'algorithme de Pavesi du programme *EufindtRNA* ([Pavesi *et al.*, 1994] Tab. 3.1 p. 79) permet de rechercher puis de combiner les signaux promoteurs et terminateurs pour prédire les ARNt (dans ce cas le repliement n'est pas étudié). L'algorithme du programme *tRNAscan* ([Fichant & Burks, 1991] Tab. 3.1 p. 79) recherche les deux promoteurs intragéniques puis utilise un système hiérarchique à base de règles, dans lequel chaque ARNt potentiel possédant les deux promoteurs doit avoir en plus la capacité de former des appariements caractéristiques des structures tige-boucle. Ce prototype d'ARNt (ou *template*[3]) peut être adapté à une espèce individuelle.

Une seconde approche met en œuvre un modèle probabiliste plus complexe (profil pour *FastRNA-CM* [el Mabrouk & Lisacek, 1996] et profil HMM pour *Cove* [Eddy & Durbin, 1994] ; voir p. 80) dérivé d'un alignement multiple des séquences complètes d'ARNt (à la différence de l'approche précédente qui utilise des alignements partiels des ARNt).

L'annotation fonctionnelle des ARNt consiste à modéliser leur structure secondaire à l'aide d'une grammaire spécifique, afin de caractériser leur anticodon [Durbin *et al.*, 2001h, Durbin *et al.*, 2001g]. Le programme *Cove* possède ces deux dernières fonctionnalités.

Bien sûr, l'idéal est de combiner les résultats des méthodes complémentaires. Ceci a été réalisé dans le cadre du programme *tRNAscan-SE* [Lowe & Eddy, 1997], qui utilise les routines des trois programmes précédemment évoqués : *tRNAscan*, *EufindtRNA* et *Cove*.

[3]Le *template* est une structure générale consensus de l'ARNt qui décrit la taille des tiges et des boucles et la présence des nucléotides conservés.

TAB. 3.1 – Exemples de programmes de recherche de motifs, de structures secondaires et tertiaires dans les séquences nucléiques procaryotes

Category	Method	Content	http	Reference
tRNA	EufindtRNA	Identification of new eukaryotic tRNA genes by a multistep weight matrix analysis of transcriptional control regions		Pavesi-1994
	tRNAscan	A general tRNA consensus structure is a template for identifying tRNA genes	bioserve@genome.lanl.gov	Fichant-1991
	Cove	RNA sequence analysis using covariance models	http://www.genetics.wustl.edu/eddy/software/#cove	Eddy-1994
	FastRNA	Fast version of tRNAscan: signal search (probabilisitic approach or pattern matching approach), and secondary structure prediction	http://www-igm.univ-mlv.fr/~mabrouk/	el-Mabrouk-1996
	tRNAscan-SE	Recherche de motifs et de structures secondaire combine trois méthodes: EufindtRNA, tRNAscan et Cove	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/	Lowe-1997
Terminator	Petrin	Recherche la structure secondaire tige-boucle des terminateurs bactériens rho-indépendants		Carafa-1990
	TransTer	Recherche la structure secondaire tige-boucle des terminateurs bactériens rho-indépendants et de la queue polyU	http://www.tigr.org/software/transterm.html	Ermolaeva-2000
Pattern Matching	Palingol	A declarative programming language to describe nucleic acids' secondary structure and to scan sequence database	http://wwwabi.snv.jussieu.fr/cgi-bin/wrap/public/Palingol/	Billoud-1996
	ADAPT	Recherche de structures tertiaires de l'ADN impliqués dans des interactions avec des protéines ou avec des ligands (champs de force)	http://www.ibpc.fr/CB2000/B_lafontaine.html	Lafontaine-2001
	Consensus	Découverte de matrices pondérées	http://ural.wustl.edu/~jhc1/consensus/	Hertz-1999
	Rmes	Découverte de mots exceptionnels (chaînes de markov)	http://www-mig.jouy.inra.fr/ssb/rmes/	Schbath-1997
	Spa	Découverte de mots exceptionnels (chaînes de markov)	http://stat.genopole.cnrs.fr/SPA/	Richard-2003
	SMILE	Découverte de patrons exceptionnels (arbre des suffixes)	http://www-igm.univ-mlv.fr/~marsan/smile.html	Marsan-2000
	FootPrinter	Découverte d'empreintes phylogénétiques à partir de séquences homologues	http://bio.cs.washington.edu/software.html	Blanchette-2003
	RepeatFinder	Découverte de répétitions exactes (arbre des suffixes)	http://www.tigr.org/software/	
	Spat	Reconnaissance exacte d'une expression régulière autorisant les erreurs (algorithme Shift-Or)	alain.viari@inria.fr	Wu-1991
	RBSfinder	Reconnaissance d'une matrice de RBS (probabilités d'occurrence de la base b à la position k) et d'un vecteur de positions (probabilités d'occurrence d'un RBS à i nucléotides en 5' du codon d'initiation) et réajustement du codon d'initiation	http://www.tigr.org/software/	Suzek-2001
	NNPP	Reconnaissance de promoteurs (réseaux de neurones)	http://www.fruitfly.org/seq_tools/promoter.html	Reese-2001
	ELPH	Recherche de motifs (Echantillonnage de Gibbs)	http://www.tigr.org/software/	
	YEBIS	Recherche de motifs (profil HMM)	http://www-btls.jst.go.jp/MotifExtraction/	Yada-98
	RSAT	Plusieurs modules d'analyse de séquences régulatrices	http://rsat.ulb.ac.be/rsat/	VanHelden-2003
	NOSFERATU	Plusieurs modules pour rechercher différents types de répétition longue	erocha@abi.snv.jussieu.fr	Achaz-2002

3.2.2 ARN ribosomique

L'annotation syntaxique et fonctionnelle des ARNr est beaucoup plus simple que celle des ARNt. En effet, puisque les séquences d'ARNr (5S, 16S et 23S) sont relativement bien conservées, il suffit de rechercher le long de la séquence d'ADN des similitudes avec des séquences d'ADNr (méthode extrinsèque). Cependant, lorsque la banque utilisée contient des séquences d'ADNr de plusieurs génomes, la compilation des résultats n'est pas forcément aisée : comment synthétiser l'information lorsque plusieurs ADNr s'alignent avec une même région chromosomique ?

Dans le programme FindrRNA développé au sein de l'AGC (Tab. 5.2 C p. 172), le problème est résolu de la façon suivante : FindrRNA utilise le programme Blast2n de la plate-forme Biofacet (matrice nucléique nuc4.4 [Glemet & Codani, 1997]) pour effectuer la comparaison de séquences entre le génome étudié et une banque d'ARNr procaryotes. Puis un autre module de cette plate-forme regroupe les fragments de séquence dont l'alignement est significatif, selon un critère donné, par exemple, la taille, les propriétés d'alignement, l'annotation fonctionnelle, etc. Dans notre cas, un groupe contient l'ensemble des fragments de séquence d'ADNr qui s'alignent avec une région génomique donnée, ainsi que le fragment génomique couvert par ces alignements.

3.3 Prédiction de gènes codant des protéines

De nombreuses méthodes intrinsèques de prédiction de gènes codant des polypeptides ont été développées pour répondre à l'émergence des nombreux projets de séquençage. La plupart de ces méthodes utilisent le fait que les CDS ont des propriétés de composition différentes des régions intergéniques. Ces programmes peuvent être divisés en deux grands groupes [Borodovsky & McIninch, 1993b] : la recherche par contenu et la recherche par signal. L'approche par contenu (ou globale) est fondée sur l'analyse statistique de fragments d'ADN suffisamment longs. Cette analyse permet de mesurer un potentiel de codage le long de la séquence afin de définir les régions codantes. La recherche par signal (ou locale) est destinée à prédire les positions des sites bornant les CDS. Il ne reste alors qu'à définir la CDS comme la région comprise entre des sites frontaliers appropriés. Nous allons aborder la recherche par signal et décrire plus précisément celle par contenu, qui est au cœur de ce travail de recherche.

3.3.1 Recherche par signal

Il existe plusieurs raisons de rechercher des motifs dans les séquences. La localisation de signaux de régulation tels que les codons d'initiation, les codons de terminaison et les sites de liaison du ribosome (RBS) permet de délimiter les régions codantes d'une séquence nucléique procaryote. La localisation d'un ou de plusieurs motifs répertoriés dans des bases de motifs protéiques peut aider à caractériser la fonction d'un polypeptide.

Définitions

On distingue le *motif*, qui est généralement un segment court, continu et non ambigu d'une séquence (site de restriction), du *patron* qui a une structure plus complexe (dégénérée et/ou composée : un consensus lié à une famille de protéines ou à un facteur de transcription). Ce dernier est souvent constitué de différents motifs plus ou moins distants les uns des autres. De plus, sa définition peut comporter des exclusions ou des associations de motifs. La difficulté majeure est de savoir quel motif utiliser et quelle est la pertinence de sa définition. La problématique de recherche de signaux dans les séquences est constituée de deux phases :

1. la phase de découverte du signal proprement dite, sa description et sa représentation
2. la phase de reconnaissance permet ensuite de rechercher les occurrences de ce signal sur une nouvelle séquence

Lors de la phase de découverte, la description d'un motif est généralement établie à partir de l'alignement d'un ensemble de séquences supposées contenir le motif (c'est l'ensemble d'apprentissage). Les jeux d'apprentissage peuvent être constitués par exemple d'éléments régulateurs tels que des opérateurs, des promoteurs, des RBS, des terminateurs (*cis-acting regulatory element* (*CARE*)) déjà connus et extraits de banques de motifs nucléiques comme ooTFD [Ghosh, 1998] ou IMD (une banque de matrices [Chen *et al.*, 1995]). Lorsqu'il est nécessaire de découvrir un motif, les objets nucléiques filtrés peuvent être de natures différentes : des régions intergéniques ou des régions en $5'$ de gènes orthologues ou corégulés. Les motifs sont de qualités diverses (courts, longs, dégénérés et/ou répétés), en quantité inégale (sous/sur-représenté), et impliqués dans des mécanismes de régulation variés (activation, répression, initiation de la transcription, de la traduction, terminaison de la transcription). Cependant, ils possèdent des caractéristiques communes : leur séquence est fondée sur un alphabet à quatre lettres (A, C, G et T), leur taille est souvent comprise entre 5 et 25 pb et ils sont présents en plusieurs exemplaires. La formalisation et la représentation de ces éléments reste parfois délicate (chaîne de caractères, mot, site, motif, signal, bloc, domaine, répétition, consensus, expression régulière, modèle, profil, matrice, patron, etc. [Bocs, 1999]) et dépendent du type de signal étudié et du type de méthode utilisée. Nous allons décrire les principales méthodes de recherche de CARE chez les bactéries, car les connaissances sur les CARE des archées sont moins avancées.

Exemples

Pour certaines bactéries le motif consensus du site *RBS* est connu dans la littérature (AGGAGG pour *B. subtilis* et [AT][CA]AGGA pour *E. coli* K-12 [Frishman *et al.*, 1998b]). Pour d'autres organismes, à partir de sites RBS caractérisés expérimentalement, un alignement multiple des séquences correspondantes est réalisé afin d'en déduire un consensus. Lorsqu'aucun RBS de l'organisme étudié n'est connu, un programme du type *SPat* ([Wu & Manber, 1991] Tab. 3.1 p. 79) permet dans

un premier temps de découvrir les RBS à partir d'un motif RBS dégénéré[4] (AGGA par exemple) dans une région d'environ 50 pb en 5′ et 3′ du codon d'initiation *présumé* des CDS. *SPat* est un programme de recherche d'expressions régulières permettant les erreurs et supportant l'alphabet dégénéré IUPAC. Un consensus est alors déduit de l'alignement multiple de RBS validés manuellement. Le motif obtenu est utilisé dans un second temps pour reconnaître les sites RBS. Certains programmes utilisent des modèles probabilistes pour détecter les RBS et aider au réajustement du codon d'initiation des CDS. Ils décrivent généralement le RBS mais aussi la distance qui le sépare du codon d'initiation. La recherche de RBS en 5′ des CDS à partir de ce modèle peut se faire de différentes manières : à l'aide d'une fenêtre glissante dans une étape de post-traitement d'un programme de prédiction de gènes [Hayes & Borodovsky, 1998a, Frishman *et al.*, 1999], dans un programme indépendant (*e.g. RBSfinder* ; [Suzek *et al.*, 2001] Tab. 3.1 p. 79), ou encore au cœur d'un programme de prédiction utilisant les modèles de Markov cachés (*Hidden Markov Model* (*HMM*) voir p. 84 [Besemer *et al.*, 2001, Larsen & Krogh, 2003, Nicolas, 2003]). Par exemple, dans le cas du modèle RBS de *GeneMark* [Hayes & Borodovsky, 1998a], la phase d'apprentissage permet de calculer les fréquences d'apparition des nucléotides à chaque position du RBS à partir d'un alignement multiple par échantillonnage de Gibbs de séquences extraites en 5′ du codon d'initiation de CDS. La phase de reconnaissance de RBS utilise une fenêtre de la taille du RBS (6 pb) glissant de la position $i = -21$ à $i = -4$ en 5′ des CDS[5]. $P(RBS_i \mid F)$, la probabilité d'être dans le modèle RBS à la position i connaissant le fragment de séquence est calculée à l'aide de la formule de Bayes qui prend en compte le modèle RBS et un modèle non RBS. La position $i*$, correspondant à la probabilité maximale, est retenue comme localisation du RBS prédit.

En ce qui concerne les signaux de régulation de la transcription, seule la recherche de *terminateurs* indépendant du facteur rho est classiquement utilisée sur un génome complet. Les programmes *Petrin* et *TransTerm* ([d'Aubenton Carafa *et al.*, 1990, Ermolaeva *et al.*, 2000] Tab. 3.1 p. 79) prédisent ce genre de signaux grâce aux calculs d'un score d'énergie (mesure de la stabilité des structures tiges-boucles) et d'un score de la queue polyU (composition et proximité de la région riche en U en 3′). De plus, *TransTerm* calcule un score de confiance qui combine ces deux scores et tient compte de la position et de l'orientation du terminateur par rapport aux gènes voisins. Ces programmes prédisent plus de 95% de terminateurs bactériens. Pour la prédiction de *promoteurs*, l'étape d'apprentissage du site de fixation de la RNA polymérase est ardue car cet élément possède une grande variabilité et n'est pas très fréquent (au moins une région promotrice par structure opéronique). A ce jour, aucun programme de prédiction de gènes procaryotes ne recherche les promoteurs. Par ailleurs, les programmes de reconnaissance de promoteurs recherchent un promoteur soit dans les régions en 5′ du +1 de transcription (cette position n'est précisément déterminée qu'à partir du séquençage de l'ADNc), soit dans les régions 5′ de gènes orthologues. Le programme *NNPP* (*Neural Network for Promoter Prediction* [Reese, 2001]) permet de rechercher les promoteurs pro-

[4]Il est aussi possible d'utiliser la séquence inverse complémentaire de l'anti-RBS de l'ARN**16S** (dans les trente derniers nucléotides).

[5]i est la position de début du RBS et -1 est la dernière base avant le codon d'initiation.

caryotes ou eucaryotes à partir d'un modèle de réseau de neurones qui combine les propriétés de structure et de composition du cœur de la région promotrice. C'est à l'heure actuelle l'un des logiciels les plus efficaces en matière de prédiction de promoteurs [Fickett & Hatzigeorgiou, 1997]. Enfin, pour la prédiction d'*opérateurs* (activation ou répression de l'initiation de la transcription), le signal recherché n'est pas toujours connu. Pour découvrir le motif, on peut rechercher des mots exceptionnels (sur-représentés) dans les régions $5'$ de gènes orthologues ou corégulés. Le module oligo-analysis du logiciel *RSAT* ([van Helden, 2003] Tab. 3.1 p. 79) compte l'occurrence de tous les oligonucléotides d'une certaine longueur, estime s'ils sont sur-représentés (fréquence observée sur la fréquence attendue) et si cette sur-représentation est statistiquement significative (calcul d'une E-value par la loi binomiale).

Méthodologie

La phase de reconnaissance de motifs correspond à une recherche de structures secondaires dans le cas des terminateurs, et à une recherche de structure primaire (séquence) dans les autres cas. Cette dernière met en œuvre des méthodes de recherche exacte d'un mot [Charras & Lecroq,] et plus généralement d'une expression régulière, ou des méthodes de reconnaissance d'un modèle de séquences (matrice de consensus [Durbin *et al.*, 2001a]). Ces deux approches nécessitent souvent de préciser la région étudiée : soit on fournit directement des fragments de séquence intergéniques, soit on définit les positions entre lesquelles s'effectue la recherche. Les méthodes de recherche exacte d'un mot sont fondées sur une recherche par signal en utilisant l'algorithme Shift-Or (*SPat*) ou l'arbre des suffixes[6] (*SMILE* [Marsan & Sagot, 2000] Tab. 3.1 p. 79). Les méthodes fondées sur des matrices sont basées sur une recherche par contenu (*Position-Specific Scoring Matrix* (*PSSM* ; *RSAT*) et permettent de donner plus d'importance à certaines positions d'un alignement multiple. Un profil est une matrice de scores spécifiques de positions consensus qui autorise des insertions-délétions pour les positions non consensus d'un alignement multiple. Un profil est donc plus adéquat qu'un PSSM pour représenter toute l'information d'un alignement multiple d'une famille de séquences protéiques [Durbin *et al.*, 2001f]. L'échantillonnage de Gibbs (PSSM ; *ELPH* Tab. 3.1 p. 79) et les HMM (profils probabilistes avec modélisation d'états d'insertion et de délétion ; *YEBIS* [Yada *et al.*, 1998]) sont deux méthodes qui permettent à la fois de rechercher et de découvrir les motifs à partir de séquences non alignées grâce à des variantes de l'algorithme *Expectation Maximisation* (*EM* voir p. 105). Enfin, une approche en voie de développement consiste à déterminer les relations entre la séquence et la structure tertiaire de l'ADN pour chercher des déformations de l'ADN et des interactions moléculaires (ADN-protéine et ADN-ligand) qui sont énergétiquement favorables (ADAPT [Lafontaine & Lavery, 2000] Tab. 3.1 p. 79).

Afin d'optimiser l'étape d'apprentissage d'un motif il faudrait disposer, dans l'idéal, de jeux de séquences caractérisés expérimentalement pour chaque espèce procaryote étudiée : promoteurs, opérateurs, RBS et terminateurs. Cette lacune est toutefois en partie comblée par les méthodes

[6] C'est un type d'automate fini.

bioinformatiques aujourd'hui disponibles qui sont de plus en plus nombreuses, efficaces et complémentaires. On espère, à court terme, découvrir des empreintes phylogénétiques[7] (*FootPrinter* [Blanchette & Tompa, 2003] TAB. 3.1 p. 79), à moyen terme, utiliser la méthode des HMM et à long terme, modéliser les structures tertiaires de l'ADN à partir de la séquence. Aujourd'hui, on pourrait imaginer une façon de combiner les résultats de différentes méthodes de recherche de signaux au sein d'un programme de prédiction de gènes pour aider à définir le codon d'initiation le plus probable des CDS, et les structures en opéron.

3.3.2 Recherche par contenu

Habituellement, pour l'identification de gènes codants des polypeptides dans les génomes procaryotes, la recherche par contenu donne de meilleurs résultats que la recherche par signal car les CDS ne sont pas interrompues par des introns (pas de signaux d'épissage). On peut diviser les méthodes de recherche par contenu en deux grands groupes.

Les *méthodes indépendantes d'un modèle d'ADN codant* reposent généralement sur l'hypothèse que l'ADN codant est moins aléatoire que l'ADN non-codant. Ces méthodes ne nécessitent pas de phase d'apprentissage. L'écart par rapport à l'aléatoire est mesuré indépendamment d'un modèle de référence et le score qui en résulte est corrélé au potentiel de codage. En 1982, Fickett et Tung proposent une méthode fondée sur l'asymétrie dans la composition en base entre les trois positions d'un codon (*TestCode* [Fickett, 1982] TAB. 3.3 p. 91). Pour déterminer si une région d'ADN est codante, on peut rechercher les irrégularités de période 3 dans la distribution des nucléotides. Dans une séquence codante, on s'attend à ce que l'utilisation non uniforme des codons se traduise par un biais de période 3 dans les fréquences individuelles d'apparition des nucléotides. On calcule pour chaque base sa fréquence d'apparition aux positions 3j, 3j+1 et 3j+2 et on la compare à sa fréquence moyenne d'apparition dans la séquence. Autrement dit, l'asymétrie d'un nucléotide est définie comme la variance de la fréquence du nucléotide aux trois positions des codons de la séquence. *TestCode* calcule l'asymétrie indépendamment pour chacun des quatre nucléotides et somme ces quatre valeurs en un unique score d'asymétrie de position de la séquence X, $PA(X)$[8]. Il existe d'autres types de méthodes indépendantes d'un modèle codant, telles que celles fondées sur les corrélations périodiques entre les positions des nucléotides (index d'asymétrie périodique, information mutuelle moyenne, spectre de fourier ; [Guigo, 1999]). Elles sont généralement moins puissantes que les méthodes dépendantes d'un modèle d'ADN codant pour discriminer l'ADN codant de l'ADN non-codant, car le potentiel de codage dépendant d'un modèle est vraisemblablement plus apte à décrire les caractéristiques spécifiques de l'ADN codant. En contrepartie, il devient nécessaire d'avoir un échantillon représentatif de l'ADN codant pour chaque espèce étudiée afin d'estimer cor-

[7] La recherche d'un élément régulateur dans des régions homologues de différentes espèces est peut-être préférable à celle d'éléments régulateurs similaires dans une même espèce (orthologue vs corégulé).

[8] On a donc $a \in \{A, C, G, T\}$, $\mu(a) = \frac{\sum_{m=1}^{3} F_R(a,pos)}{3}$; $asym(a) = \sum_{pos=1}^{3}(F_R(a,pos) - \mu(a))^2$; $PA(X) = asym(A) + asym(C) + asym(G) + asym(T)$, où $F_R(a,pos)$ est la fréquence relative du nucléotide a à la position pos du codon, calculée à partir d'une des trois décompositions de la séquence X en codons (n'importe laquelle).

rectement les paramètres du modèle. Plus le modèle est complexe et plus il est sensible à la taille de l'échantillon et aux biais d'échantillonnage.

Les *méthodes dépendantes d'un modèle d'ADN codant* reposent sur un modèle probabiliste et exigent une phase d'apprentissage manuelle ou automatique. Ces méthodes nécessitent tout d'abord de choisir le modèle. On estime alors les paramètres du modèle par maximum de vraisemblance sur un grand jeu de CDS de l'espèce procaryote (phase d'apprentissage). Puis, on calcule le potentiel de codage d'une nouvelle séquence (phase de reconnaissance), de deux manières. Soit on applique le calcul du potentiel de codage directement aux CDS complètes, ce qui génère pour chaque CDS six probabilités de codage (une pour chaque phase). Soit on applique ce même calcul à un fragment nucléique contenu dans une fenêtre glissant le long de la séquence, afin d'obtenir six probabilités de codage à chaque position de la fenêtre sur la séquence, utilisées alors pour représenter graphiquement des courbes de prédiction de codage le long des six phases de lecture de la séquence.

Les méthodes de recherche par contenu reposent sur le calcul d'un potentiel de codage encore appelé statistique de codage. Le potentiel de codage d'une séquence d'ADN donnée est défini par une fonction qui calcule un nombre réel lié à la vraisemblance que cette séquence code un polypeptide. Avant d'entrer dans la description des principaux programmes de prédiction de gènes dépendant d'un modèle d'ADN codant, rappelons quelques éléments de statistique.

Modèles probabilistes de séquences d'ADN et statistiques

Un modèle mathématique est un objet mathématique substitué à la réalité dans un but déterminé. Le calcul des probabilités permet de modéliser la variabilité. Dans un modèle probabiliste, on considère que les données observées sont des réalisations de variables aléatoires. Un modèle probabiliste de séquences permet de calculer la probabilité d'apparition de chaque séquence. On peut séparer ces modèles en deux groupes : les modèles indépendants (modèle de permutation, modèle de Bernoulli) et les modèles à dépendance (modèle de chaînes de Markov, arbre de contexte, modèle de chaînes de Markov cachées, modèle semi-markovien caché). Lors de la confrontation de modèles probabilistes avec la réalité, se posent les problèmes de l'estimation des paramètres et de l'évaluation de la significativité des résultats par rapport à un modèle aléatoire : on parle alors de modèles statistiques. Les séquences d'ADN peuvent être représentées comme une succession de lettres choisies dans l'alphabet à quatre lettres $\mathcal{A} = \{A, C, G, T\}$. Soit une séquence de nucléotides de longueur n : $X_1^n = X_1 \dots X_i \dots X_n,\ X_i \in \mathcal{A}$ et soit la séquence de triplets de longueur l : $X_1^l = X_1 Y_1 Z_1 \dots X_j Y_j Z_j \dots X_l Y_l Z_l,\ X_j, Y_j, Z_j \in \mathcal{A},\ l = n/3$.

Modèle de Bernoulli ($M0$) Le modèle de Bernoulli fait l'hypothèse d'indépendance d'apparition des nucléotides à chacune des positions de la séquence. Les paramètres de ce modèle, π, correspondent aux probabilités d'apparition des nucléotides dans la séquence $\pi(A)$, $\pi(C)$, $\pi(G)$ et $\pi(T)$:

$$\pi(a) = \mathbb{P}_\pi(X_i = a),\ a \in \mathcal{A}.$$

Les contraintes sur ces paramètres sont $\pi(a) \in [0,1]$ et $\sum_{a\in\mathcal{A}}\ \pi(a) = 1$. La probabilité d'apparition d'une séquence sous ce modèle est :

$$\begin{aligned}\mathbb{P}_\pi(X_1^n = x_1^n) &= \mathbb{P}_\pi(X_1 = x_1)\dots\mathbb{P}_\pi(X_i = x_i)\dots\mathbb{P}_\pi(X_n = x_n),\ x_i \in \mathcal{A}\\ &= \pi(x_1)\dots\pi(x_i)\dots\pi(x_n)\\ &= \prod_{a\in\mathcal{A}}(\pi(a))^{N_a}.\end{aligned}$$

N_a est le nombre d'occurrence de a dans la séquence. La probabilité d'apparition de la séquence ne dépend que de sa composition en nucléotides.

En pratique, pour utiliser ce modèle, il faut choisir une valeur pour les paramètres. Généralement, on utilise l'estimateur du maximum de vraisemblance : $\hat{\pi}(a) = N_a/n$. Autrement dit, on maximise $\mathbb{P}_\pi(X_1^n)$ par les fréquences observées dans un échantillon de séquences biologiques représentatif du modèle (jeu d'apprentissage).

Chaîne de Markov (Mm) Le modèle Mm part de l'hypothèse que la probabilité d'apparition d'un nucléotide dépend des m nucléotides qui précèdent, ce qui permet de prendre en compte la composition en oligonucléotides (mots) de longueur $m+1$ dans les séquences d'ADN (Fig. 3.1 p. 87). Les paramètres markoviens correspondent aux probabilités de transition, π, et aux probabilités initiales, μ. La structure de dépendance d'ordre m est donnée par la matrice de transition de la chaîne de Markov Π de dimension $\mid \mathcal{A} \mid^m * \mid \mathcal{A} \mid$. La valeur $\mathbb{P}_\pi(X_i = b \mid X_{i-m} = a_1, X_{i-m+1} = a_2, \dots, X_{i-1} = a_m)$ est donnée par l'élément noté $\pi(a_1\dots a_m, b)$ de cette matrice de transition :

$$\pi(a_1\dots a_m, b) = \mathbb{P}_\pi(X_i = b \mid X_{i-m} = a_1, X_{i-m+1} = a_2, \dots, X_{i-1} = a_m),\ a, b \in \mathcal{A}.$$

Les sommes en ligne de la matrice de transition font 1 : $\sum_{b\in\mathcal{A}}\ \pi(a_1\dots a_m, b) = 1, \forall a_1, \dots, a_m \in \mathcal{A}$. Il existe $4^m * (4-1)$ paramètres de transition. Les probabilités initiales correspondent à :

$$\mu(a_1\dots a_m) = \mathbb{P}_\pi(X_1 = a_1, \dots, X_m = a_m),\ a \in \mathcal{A}.$$

La probabilité d'apparition d'une séquence sous le modèle $M1$ est :

$$\begin{aligned}\mathbb{P}_\pi(X_1^n = x_1^n) &= \mathbb{P}_\pi(X_1 = x_1)\mathbb{P}_\pi(X_2 = x_2 \mid X_1 = x_1)\dots\mathbb{P}_\pi(X_i = x_i \mid X_{i-1} = x_{i-1})\dots\\ &\quad \mathbb{P}_\pi(X_n = x_n \mid X_{n-1} = x_{n-1}),\ x_i \in \mathcal{A}\\ &= \mu(x_1)\pi(x_1, x_2)\dots\pi(x_{i-1}, x_i)\dots\pi(x_{n-1}, x_n)\\ &= \mu(x_1)\prod_{a,b\in\mathcal{A}}(\pi(a,b))^{N_{ab}}.\end{aligned}$$

La probabilité d'apparition d'une séquence sous le modèle Mm est :

$$\begin{aligned}\mathbb{P}_\pi(X_1^n = x_1^n) &= \mu(x_1\dots x_m)\pi(x_1\dots x_m, x_{m+1})\dots\pi(x_{n-m}\dots x_{n-1}, x_n)\\ &= \mu(x_1\dots x_m)\prod_{a_1\dots a_m, b\in\mathcal{A}}(\pi(a_1\dots a_m, b))^{N_{a_1\dots a_m b}}.\end{aligned}$$

La probabilité d'apparition d'une séquence ne dépend que de sa composition en mots de longueur $m+1$.

On estime les paramètres selon le même principe de maximum de vraisemblance :

$$\begin{aligned}\hat{\pi}(a_1 \dots a_m, b) &= \frac{N_{a_1 \dots a_m b}}{N_{a_1 \dots a_m}} \\ \hat{\mu}(a_1 \dots a_m) &= \frac{N_{a_1 \dots a_m}}{n-m+1}\end{aligned}$$

.

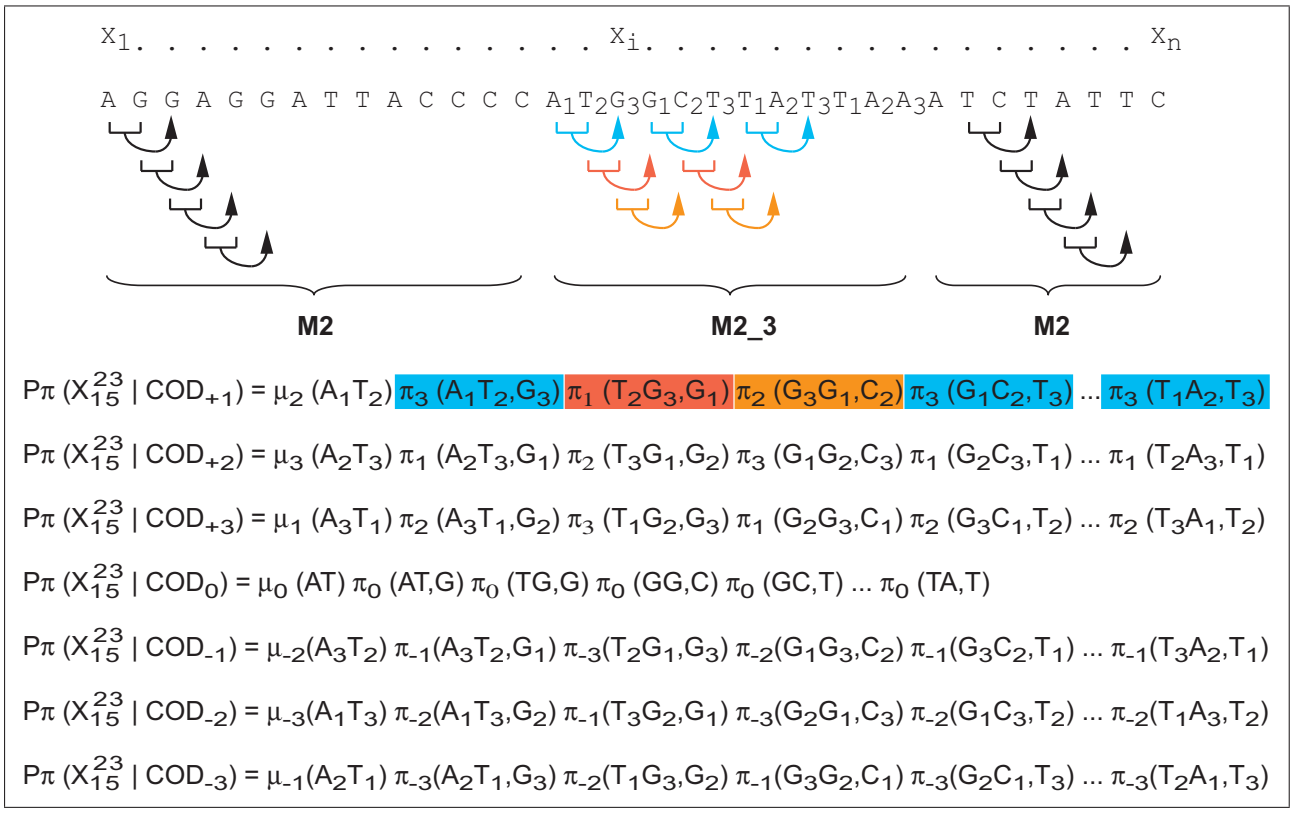


FIG. 3.1 – Chaînes de Markov pour modéliser les séquences d'ADN
Sur cet exemple, la séquence est codante en phase +1 de la position 15 (ATG) à la position 26 (TAA). La fréquence des nucléotides s'ajuste sur les mots de longueur 3 et dépend de la position dans le codon ($M2\_3$). Ce modèle permet de calculer la probabilité d'apparition d'un fragment sachant qu'il est codant en phase +1, +2, +3, -1, -2, -3 ou qu'il est non-codant (phase 0).

Chaîne de Markov à transitions 3-périodique (Mm_3) Les chaînes de Markov d'ordre m vu précédemment sont homogènes le long de la séquence, c'est-à-dire que les probabilités $\mathbb{P}(X_i = b \mid X_{i-m} = a_1, \dots, X_{i-1} = a_m)$ et $\mathbb{P}(X_{i-m} = a_1, \dots, X_{i-1} = a_m)$ ne dépendent pas de la position i dans la chaîne X_1^n.

Dans les régions codantes, génétiquement découpées en triplets adjacents (codons), la fréquence d'apparition d'un nucléotide varie aussi selon sa position dans le codon (FIG. 3.1 p. 87). Il est bien

sûr très intéressant, notamment en vue de retrouver la phase de lecture, de prendre en compte cette périodicité. La modélisation de la périodicité de composition des CDS nécessite d'introduire une loi d'apparition des nucléotides différente pour chacune des trois positions des codons. Les paramètres de ce modèle se séparent en trois jeux Π_1, Π_2 et Π_3, où Π_{pos} décrit la fréquence d'apparition des nucléotides en position *pos* des codons. Sous le modèle $M1$, La probabilité de transition de a vers b va alors dépendre de la position de la lettre b dans la séquence, et trois probabilités de transition sont ainsi différenciées :

$$\begin{aligned}
\pi_1(a,b) &= \mathbb{P}_\pi(X_{3j+1} = b \mid X_{3j} = a),\ a,b \in \mathcal{A},\ \forall j \\
\pi_2(a,b) &= \mathbb{P}_\pi(X_{3j+2} = b \mid X_{3j+1} = a) \\
\pi_3(a,b) &= \mathbb{P}_\pi(X_{3j+3} = b \mid X_{3j+2} = a).
\end{aligned}$$

On suppose ici que la séquence commence par un codon *entier*. Dans un tel modèle, il existe donc trois fois plus de paramètres que dans un modèle $M1$. De la même façon, il y a trois lois initiales, μ_1, μ_2 et μ_3 suivant la position du nucléotide :

$$\begin{aligned}
\mu_1(a) &= \mathbb{P}_\pi(X_{3j+1} = a),\ a \in \mathcal{A},\ \forall j \\
\mu_2(a) &= \mathbb{P}_\pi(X_{3j+2} = a) \\
\mu_3(a) &= \mathbb{P}_\pi(X_{3j+3} = a).
\end{aligned}$$

Sous ce modèle noté $M1\_3$ et dit 3-périodique (ou phasé), la probabilité d'apparition de la séquence x_1^n s'écrit de trois façons différentes selon la phase COD_1, COD_2 ou COD_3 dans laquelle la séquence code :

$$\begin{aligned}
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_1) &= \mu_1(x_1)\pi_2(x_1,x_2)\pi_3(x_2,x_3)\pi_1(x_3,x_4)\dots \\
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_2) &= \mu_2(x_1)\pi_3(x_1,x_2)\pi_1(x_2,x_3)\pi_2(x_3,x_4)\dots \\
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_3) &= \mu_3(x_1)\pi_1(x_1,x_2)\pi_2(x_2,x_3)\pi_3(x_3,x_4)\dots
\end{aligned}$$

On estime les paramètres par maximum de vraisemblance (la position d'un oligonucléotide est défini par la position de son dernier nucléotide dans le codon) :

$$\begin{aligned}
\hat{\pi}_{pos}(a,b) &= \frac{N_{ab,pos}}{N_{a,pos-1}} \\
\hat{\mu}_{pos}(a) &= \frac{N_{a,pos}}{l}.
\end{aligned}$$

$N_{ab,pos}$ est le nombre d'occurrences du dinucléotide ab en position *pos*.

Le modèle $M2\_3$ est celui qui s'ajuste, entre autre, sur la fréquence des trinucléotides en position 3, *i.e.* la fréquence des codons. Utiliser ce modèle dans l'analyse d'une séquence codante permet ainsi de prendre en compte un éventuel biais des codons et notamment l'absence de codon de stop en phase : $\pi_3(tg,a) = 0$, $\pi_3(ta,a) = 0$ et $\pi_3(ta,g) = 0$.

$$\begin{aligned}
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_1) &= \mu_2(x_1x_2)\pi_3(x_1x_2,x_3)\pi_1(x_2x_3,x_4)\pi_2(x_3x_4,x_5)\dots \\
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_2) &= \mu_3(x_1x_2)\pi_1(x_1x_2,x_3)\pi_2(x_2x_3,x_4)\pi_3(x_3x_4,x_5)\dots \\
\mathbb{P}_\pi(X_1^n = x_1^n \mid COD_3) &= \mu_1(x_1x_2)\pi_2(x_1x_2,x_3)\pi_3(x_2x_3,x_4)\pi_1(x_3x_4,x_5)\dots
\end{aligned}$$

On modélise de la même façon les CDS sur le brin complémentaire dont on notera les phases de lecture COD_{-1}, COD_{-2} ou COD_{-3}. La probabilité d'apparition d'une séquence sous le modèle Mm_3 ne dépend que de sa composition en mots de longueur $m+1$ en position 1, 2 et 3 des codons.

Nous allons décrire les principaux modèles d'ADN codant tels que ceux fondés sur des comptages d'oligonucléotides et sur des dépendances entre les positions des nucléotides. Cependant, il existe aussi des modèles qui reposent sur des biais de compositions en bases entre les trois positions du codon (*Codon Prototype* ; [Fickett & Tung, 1992]).

Méthodes de prédiction de gènes fondées sur les tables d'usage des codons

Certaines méthodes de prédiction de gènes modélisent le jeu des CDS de l'ensemble d'apprentissage à partir de comptage d'oligonucléotides représentés par des tables d'usage des codons (ou d'usage d'hexamères [Guigo, 1999]).

		Somme des 61 valeurs pour une CDS
Fréquence absolue	$F(abc) = N_{abc}$	ℓ
Fréquence relative	$F_R(abc) = \frac{N_{abc}}{\ell}$	1
Fréquence relative des codons synonymes	$F_{RS}(abc) = \frac{N_{abc}}{\Sigma N_{a'b'c'}}$	20
Usage relatif des codons synonymes	$RSCU(abc) = \frac{N_{abc}}{\Sigma N_{a'b'c'}} * N_{syn_{abc}}$	61

TAB. 3.2 – Différentes mesures de fréquences de codons à ne pas confondre
N_{abc}, nombre d'occurrences du codon *abc* dans une CDS ; $a, b, c \in \mathcal{A}$; l, nombre de codons dans la CDS ; $\sum N_{a'b'c'}$, somme des nombres d'occurrences des codons synonymes à *abc* (autrement dit, c'est le nombre d'occurrences de l'acide aminé encodé par *abc* dans la CDS) ; $N_{syn_{abc}}$, nombre de codons synonymes au codon *abc*.

Programme Orpheus Le programme de prédiction de gènes *Orpheus* ([Frishman *et al.*, 1998b], TAB. 3.3 p. 91) modélise l'usage du code génétique. D. Frishman part du principe que les méthodes intrinsèques et extrinsèques prises séparément, ne peuvent assurer une prédiction correcte. Il préconise donc de combiner autant d'évidences que possible, afin d'obtenir des résultats de confiance. Le concept clé de cette méthode est la *graine* de CDS : c'est une CDS de confiance, de longueur minimale, inférée de manière extrinsèque (E) ou intrinsèque (I).

La phase d'apprentissage du programme *Orpheus* est fondée sur une méthode extrinsèque qui permet d'extraire un jeu de *graines* E à partir des résultats de similitude avec les séquences pro-

téiques d'une banque non redondante. Une *graine E* est obtenue en étendant une région alignée de confiance en 5′ jusqu'au premier codon d'initiation rencontré et en 3′ jusqu'au premier codon de terminaison. Ce jeu de *graines E* est utilisé pour construire la table de fréquences relatives des 61 codons de l'organisme étudié. La fréquence relative d'un codon ($F_R(abc)$) est calculée à partir de sa fréquence absolue ($F(abc)$; TAB. 3.2 p. 89).

La phase de reconnaissance d'*Orpheus* est fondée sur une méthode intrinsèque qui consiste à mesurer le potentiel de codage des régions génomiques non couvertes par le jeu de *graines E*. Le potentiel de codage, appelé qualité de codage, est fondé sur le log-vraisemblance de la fréquence relative des codons et permet de supprimer l'hétérogénéité de longueur des CDS, l'influence de la composition locale en bases et les ombres des CDS (voir p. 93). Une *graine I* est obtenue en étendant, comme précédemment, une région de longueur supérieure à 300 pb et de qualité de codage supérieure à -1 (valeur seuil par défaut pour le jeu de *graines I*).

Une phase finale de post-traitement permet d'affiner la position du codon d'initiation des CDS retenues (jeu de *graines E+I*) à l'aide d'un modèle RBS. Les ORF du jeu de *graines E+I* n'ayant qu'*un seul codon d'initiation* possible (lors de l'extension en 5′) et aucune CDS voisine à moins de 30 pb, sont alignées par rapport à leur codon d'initiation (le premier nucléotide de ce codon correspond à la position +1). Les séquences de -20 à $+3$ constituent le jeu d'apprentissage à partir duquel sont dérivés les paramètres du modèle RBS. Le modèle RBS est constitué d'une matrice de scores de nucléotides spécifiques de la position pour modéliser le site *Shine et Dalgarno* (*SD*) et d'un vecteur de positions de la boîte SD pour modéliser la distance entre le site SD et le codon d'initiation. Un score, dans la matrice, correspond au log du ratio de la fréquence d'apparition d'un nucléotide à une position donnée du site SD, normalisée par celle du nucléotide le plus probable pour cette position (*G* pour la position 5 du motif RBS d'*E. coli* K-12 et de *B. subtilis*). Un score, dans le vecteur, correspond au log du ratio du comptage de sites SD à une position donnée de la région RBS, normalisé par celui de la position la plus fréquente (-13 dans le cas d'*E. coli* K-12 et de *B. subtilis*). La phase d'apprentissage consiste en une procédure en deux temps. Dans un premier temps, la matrice poids-positions est calculée par un processus itératif en deux étapes qui permet une ré-estimation successive des paramètres jusqu'à convergence :

1. trouver dans chaque séquence le fragment de longueur L ayant le plus haut score de site SD
2. recalculer la matrice de poids-positions

Le score d'un site SD (*ACGGGG*) est la somme des scores des nucléotides de la matrice. Dans un second temps, le vecteur de positions est calculé par le même type d'approche. La force d'un RBS est la somme du score du site SD et du score de position. La phase d'assignation du codon d'initiation des CDS consiste à utiliser ce modèle pour évaluer la force des RBS en 5′ des codons d'initiation alternatifs. Finalement, pour les CDS *ayant de multiples codons d'initiation*, on choisit, parmi les RBS suffisamment forts, celui qui se situe le plus en 5′.

Programme *Codon Preference* Dans le programme *Codon Preference* ([Gribskov *et al.*, 1984] TAB. 3.3 p. 91), le potentiel de codage mesure l'usage inégal des codons synonymes à partir de la

TAB. 3.3 – Prédiction de régions codantes procaryotes

Prédiction gènes	contenu	http	Référence
TestCode	Mesure du biais de composition en base entre les différentes positions d'un codon (pas de phase d'apprentissage).	http://www.accelrys.com/products/gcg_wisconsin_package/	Fickett-1982
CodonPreference	Mesure du biais dans l'usage des codons synonymes (phase d'apprentissage manuelle).	http://www.accelrys.com/products/gcg_wisconsin_package/	Gribskov-1984
Orpheus	Le potentiel de codage est basé sur la fréquence des codons de gènes obtenus par recherche de similitude.	http://pedant.gsf.de/orpheus/	Frishman-1998
GeneMark	Chaînes de Markov (phase d'apprentissage manuelle)	http://opal.biology.gatech.edu/GeneMark/	Borodovsky-1993-b
Prokov	Chaînes de Markov (phase d'apprentissage manuelle ou automatique).	alain.viari@inria.fr	Romanet-2001
GeneMark-Genesis	Apprentissage automatique pour dériver des modèles utilisés par les méthodes de prédiction de gènes (chaînes de Markov).		Hayes-1998
Programme d'Audic et Claverie	Chaînes de Markov (phase d'apprentissage automatique).		Audic-1998
Glimmer	« Gene Locator and Interpolated Markov Modeler » (phase d'apprentissage manuelle ou automatique).	http://www.tigr.org/software/glimmer/	Delcher-1999-b
FrameD	Un graphe dirigé sans circuit modélise CDS, frameshifts et recouvrements même sens. Arrêtes pondérées par éléments de base de GLIMMER (phase d'apprentissage manuelle). Programmation dynamique pour trouver un plus court chemin.	http://www.toulouse.inra.fr/FrameD/cgi-bin/FD	Schiex-2003
EcoParse	HMM sur les nucléotides, apprentissage mixte: manuel / extrinsèque et automatique (EM) / intrinsèque, et reconnaissance par Viterbi.	ecoparse@cse.ucsc.edu	Krogh-1994
EasyGene	HMM sur les nucléotides, apprentissage automatique qui nécessite une étape d'extraction de séquences et reconnaissance par EM.	http://www.cbs.dtu.dk/services/EasyGene/	Krogh-2003
Frame-by-Frame	HMM sur les triplets, apprentissage manuel et reconnaissance par Viterbi de chaque phases de lecture prise individuellement.	http://opal.biology.gatech.edu/GeneMark/	Shmatkov-1999
Show	"Structured homogeneities Watcher": HMM sur les nucléotides, apprentissage par EM automatique (non supervisé) et intrinsèque, et reconnaissance par EM.	http://www-mig.jouy.inra.fr/ssb/SHOW/	Nicolas-2003
GeneMark.hmm	HSHM sur des chaînes de nucléotides de taille variable uniquement la phase de reconnaissance par Viterbi, la phase d'apprentissage est faite par un autre programme (Besemer-1999).	http://opal.biology.gatech.edu/GeneMark/	Lukashin-1998
GeneMarkS	GeneMark.hmm et échantillonnage de Gibbs, itératifs pour un apprentissage automatique (non supervisé).	http://opal.biology.gatech.edu/GeneMark/	Borodovsky-2001
Rhom	Recherche de régions homogènes dans une séquence: ensemble de programmes (HMM et EM) pour segmenter une séquence d'ADN (apprentissage automatique).	http://www-mig.jouy.inra.fr/ssb/rhom/	Nicolas-2002

table des fréquences relatives des codons synonymes (F_{RS}) des 61 codons dans les CDS. La fréquence relative des codons synonymes utilisée ici permet de s'affranchir d'un éventuel biais dans l'usage des acides aminés. Le modèle d'usage des codons synonymes est conditionné par la séquence protéique observée. Autrement dit, l'ensemble des séquences dont on évalue la probabilité d'apparition code la même séquence protéique[9]. La fréquence relative du codon *GGG* codant l'acide aminé *Gly* : $F_{RS}(GGG) = (N_{GGG})/(N_{GGG} + N_{GGA} + N_{GGT} + N_{GGC})$ (Tab. 3.2 p. 89). Ainsi, le calcul de F_{RS} permet de normaliser la fréquence des codons par rapport au nombre d'occurrences en acide aminé (la somme des F_{RS} des codons codant le même acide aminé vaut 1).

La première étape du programme *Codon Preference* consiste à construire la table d'usage des codons synonymes en calculant les 61 valeurs F_{RS} sur l'ensemble des CDS du jeu d'apprentissage.

Dans une seconde étape, ou phase de reconnaissance, *Codon Preference* parcourt les trois phases de lecture d'un des brins d'une nouvelle séquence et mesure si les valeurs F_{RS} observées s'apparentent à celles qui ont été calculées dans la table d'usage des codons synonymes. On déplace une fenêtre sur la séquence dont la taille est un multiple de trois. La probabilité de codage du fragment contenu dans la fenêtre est calculée, pour les trois phases de lecture, de la façon suivante :

$$\begin{aligned}
\mathbb{P}_\pi(X_1^l = x_1^l \mid COD_1) &= F_{RS}(x_1y_1z_1)F_{RS}(x_2y_2z_2)F_{RS}(x_3y_3z_3)\dots F_{RS}(x_ly_lz_l)\\
\mathbb{P}_\pi(X_1^l = x_1^l \mid COD_2) &= F_{RS}(y_1z_1x_2)F_{RS}(y_2z_2x_3)F_{RS}(y_3z_3x4)\dots F_{RS}(y_{l-1}z_{l-1}x_l)\\
\mathbb{P}_\pi(X_1^l = x_1^l \mid COD_3) &= F_{RS}(z_1x_2y_2)F_{RS}(z_2x_3y_3)F_{RS}(z_3x4y4z4)\dots F_{RS}(z_{l-1}x_ly_l).
\end{aligned}$$

La probabilité d'apparition de la séquence, sachant qu'elle est traduite en phase f et que l'on ne tient pas compte des biais en acides aminés, est d'autant plus forte que la distribution des triplets *ressemble* à celle des codons synonymes de la table utilisée. Il est possible de superposer ces probabilités de codage sur chacune des trois phases du brin direct de la séquence. Dans cette représentation graphique, les pics mettent en évidence les phases codantes de la séquence génomique.

Parmi les méthodes de prédiction de CDS qui reposent sur des comptages d'oligonucléotides : usage du code génétique ($F_R(abc)$), usage des acides aminés (table des fréquences relatives des 20 acides aminés des CDS traduites $F_{RA}(abc)$) et usage des codons synonymes ($F_{RS}(abc)$), la première est la plus discriminante. En effet, le biais d'usage du code génétique est la somme du biais d'usage des acides aminés et du biais d'usage des codons synonymes [Guigo, 1999]. Ces méthodes prennent en compte les corrélations entre nucléotides adjacents uniquement au sein d'un codon.

Méthodes fondées sur les modèles de chaînes de Markov

D'autres méthodes de prédiction de gènes sont fondées sur la dépendance entre les positions des nucléotides. L'utilisation de modèles de chaînes de Markov, qui tiennent compte de la composition en mots des séquences, est particulièrement adaptée à l'analyse des séquences biologiques. Les chaînes de Markov permettent de modéliser et de révéler des corrélations persistantes entre nucléotides adjacents.

[9] $P(X, C) = P(X \mid C) * P(C)$

Programme *GeneMark* *GeneMark* ([Borodovsky & McIninch, 1993a] TAB. 3.3 p. 91) est un programme de reconnaissance de gènes fondé sur la modélisation des séquences nucléiques à l'aide des chaînes de Markov et qui fonctionne en parallèle, sur les deux brins d'ADN. Par exemple, dans le cas d'une chaîne de Markov d'ordre 1 ($M1$), la loi de probabilité d'apparition d'une lettre en un site dépend uniquement de la lettre présente au site précédent. Les régions codant des protéines et non-codantes sont considérées comme des séquences nucléotidiques ayant des règles d'ordonnancement des nucléotides différentes, sélectionnées par les processus d'évolution de l'expression génique. La méthode *GeneMark*, dont la première version a été développée en 1986 par M. Borodovsky, présente les caractéristiques suivantes :

1. Les régions fonctionnellement différentes (séquences d'ADN non-codantes et codantes) sont représentées par des modèles de chaînes de Markov de types différents (respectivement homogène et périodique). La phase d'apprentissage est manuelle et nécessite un jeu de CDS et un jeu de séquences non-codantes pour estimer les paramètres des modèles. La phase de reconnaissance consiste à utiliser ces modèles pour prédire les phases codantes d'une nouvelle séquence d'ADN.
2. Le potentiel de codage d'un fragment donné de la séquence est défini par la probabilité *a posteriori* d'un modèle connaissant le fragment, calculée à partir du théorème de Bayes.
3. Le potentiel de codage est mesuré à l'aide d'une fenêtre glissant le long de la séquence (au lieu de considérer les CDS). Ce choix est lié au fait que M. Borodovsky avait l'intention de généraliser cette approche aux séquences d'ADN eucaryotes sur lesquels les CDS sont morcelées en exons. Dans le cas des séquences procaryotes, il est possible de faire de la prédiction de gènes en associant à chaque CDS une probabilité de codage, alors que dans le cas des eucaryotes, toute la difficulté consiste à définir les positions des exons, la prédiction de codage étant l'un des critères pouvant aider à les définir. Deux autres raisons peuvent aussi motiver ce choix :
   – L'approche par fenêtre glissante permet une mesure locale des probabilités *a posteriori*, il est donc préférable qu'elle soit la plus petite possible tout en contenant suffisamment d'information. Autrement dit, il faut trouver un compromis entre un modèle réaliste où pour que le fragment soit complètement codant en phase f ou complètement non-codant il vaut mieux avoir une petite fenêtre, et une mesure significative (il est préférable d'avoir une fenêtre suffisamment grande (> 60 pb) [Nicolas, 2003]). En revanche, l'approche qui mesure les probabilités *a posteriori* sur la CDS ne pose pas le problème du choix de la taille de la fenêtre puisqu'elle revient à avoir une fenêtre dont la taille est ajustée à celle de la CDS. Cependant, cette seconde approche est moins réaliste, notamment dans le cas des « fausses grandes » CDS (voir p. 246). En effet, décider si une CDS est « vraie ou fausse » n'est pas une décision statistique facile à prendre, surtout dans le cas des « fausses » CDS qui peuvent être un mélange de régions intergéniques et de régions codantes dans les autres phases (plus la « fausse » CDS est longue et plus le problème est accentué ; annexe p. 389 ; [Nicolas, 2003]).

Autrement dit, plus la « fausse » CDS est grande, plus la probabilité de trouver des codons de terminaison dans les autres phases et donc de la prédire à tort comme « vraie » CDS est importante (faux-positif). Les valeurs calculées par cette approche sont généralement plus élevées que la probabilité moyenne de codage calculée à partir de l'approche par fenêtre glissante. Finalement, les valeurs calculées par cette dernière approche sont influencées par la taille de la fenêtre (et le pas), mais pas par celle de la CDS[10].

– Le cas de figure extrême d'un décalage du cadre de lecture présentant une forte probabilité de codage en absence de CDS, ne sera pas vue par l'approche considérant directement les CDS (voir p. 119).

4. Les deux brins d'ADN sont analysés simultanément. Initialement, les méthodes de prédictions de gènes nécessitaient de lancer l'analyse une fois sur le brin direct, puis une fois sur le brin inverse. Quelque soit la méthode de prédiction de gènes sous jacente, cette stratégie génère de « faux » signaux de prédiction pour le brin analysé, la « vraie » région codante pouvant être localisée sur le brin complémentaire. L'origine de ce bruit est liée à l'observation bien connue que les régions codantes comportent un excès de codons de type RNY[11] [Shepherd, 1981]. Puisque cette formule est auto-complémentaire, il n'est pas étonnant de trouver des triplets RNY dans le fragment de séquence inverse complémentaire du « vraie » gène [12]. Ce « faux » signal est appelé *ombre* ou *empreinte* de la CDS (séquence inverse complémentaire de la CDS). Le formalisme de Bayes est très flexible, car il permet d'élargir le nombre de situations possibles qui doivent être distinguées les unes des autres. Aussi, en ajoutant un modèle de chaînes Markov non homogène pour les séquences d'*empreinte*, il devient possible de s'affranchir du problème général des « fausses » prédictions dans les régions d'ombre des « vrais » gènes et d'analyser les deux brins simultanément. En fait, rechercher les séquences ombres du codant sur le brin direct revient à rechercher les CDS sur le brin inverse ; c'est pourquoi on considère par la suite que seul le brin direct doit être analysé en utilisant trois types de modèle : les CDS sur le brin direct, les CDS sur le brin inverse (ombres du codant) et les séquences non-codantes.

Le principe de la méthode *GeneMark* détaillée en annexe (annexe C p. 389) peut être résumé comme suit :

Le programme *Mkmat* est dédié à la *phase d'apprentissage* de la méthode de prédiction de gènes par chaînes de Markov. l'ordre m des modèles est choisi en fonction de la taille des jeux d'apprentissage (ensemble de séquences codantes et non-codantes). Grossièrement, on peut donner une approximation de N, la taille minimum du jeu d'apprentissage nécessaire à une estimation fiable des paramètres, en considérant qu'il faut en moyenne 100 effectifs par $m+1$-uple : $N \geq 100 * 4^{(m+1)}$

[10] Les deux approches prédisent mal les très petites CDS (l<63pb) voir p. 197.

[11] R est une purine (G ou A) dans l'alphabet dégénéré IUPAC et Y une pyrimidine (C ou T).

[12] Chez *E. coli* K-12, les « faux » signaux sont plus importants lorsque l'on se trouve dans une région du génome contenant des gènes fortement exprimés. Une explication possible serait liée au fait que l'usage des codons synonymes de ces gènes est généralement décalé vers les codons optimaux dont certains sont de la forme RNY [Borodovsky & McIninch, 1993a].

(voir p. 40). Plus précisément, le manuel de *Mkmat*[13] conseille d'utiliser les équations suivantes :

$$\begin{aligned} \sum N &\geq 30 * 4^{(m+1)} \\ \sum N &\geq 90 * 4^{(m+1)}, \end{aligned}$$

pour donner une approximation de N dans le cas des modèles, respectivement, non-codant et codant. Il faut donc trois fois moins de séquence pour estimer correctement les fréquences d'une chaîne de Markov homogène que d'une chaîne de Markov tri-périodique. Par exemple, dans le cas d'un chromosome de plus de 4000 kb, il est raisonnable d'utiliser un ordre 5, car cela revient à compter les hexamères sur des échantillons d'au moins 123 et 370 kb, respectivement pour le non-codant et le codant. La phase d'apprentissage du programme *MkMat* nécessite un jeu de CDS et un jeu de séquences non-codantes, de préférence *natives* (régions intergéniques et gènes spécifiant des ARN). Les paramètres des modèles sont estimés par maximum de vraisemblance aux trois positions des codons pour le jeu de séquences codantes, et indépendamment de la position dans le codon pour le jeu de séquences non-codantes.

Le programme *GeneMark discrimine*, sur la séquence nucléique analysée, les phases codantes des régions non-codantes, en référence aux paramètres des modèles codant et non-codant, estimés et stockés dans une matrice par *Mkmat* : c'est la *phase de reconnaissance* de la méthode de prédiction de gènes. Pour cela, *GeneMark* calcule :

- les trois vraisemblances $\mathbb{P}_\pi(X_1^w = x_1^w \mid COD_f),\ f \in \{1,2,3\}$ suivant les trois phases de lecture d'un fragment de séquence pour le modèle codant ;
- les trois vraisemblances $\mathbb{P}_\pi(X_1^w = x_1^w \mid COD_f),\ f \in \{-1,-2,-3\}$ suivant les trois phases de lecture pour le modèle ombre du codant ;
- la vraisemblance pour le non-codant $\mathbb{P}_\pi(X_1^w = x_1^w \mid COD_0)$.

A partir de ces sept valeurs et des valeurs de probabilités *a priori*, la formule de Bayes permet de calculer les sept valeurs de probabilités *a posteriori* :

- $\mathbb{P}_\pi(COD_f \mid X_1^w = x_1^w),\ f \in \{1,2,3\}$ pour les trois phases de lecture du fragment dans le modèle codant direct
- $\mathbb{P}_\pi(COD_f \mid X_1^w = x_1^w),\ f \in \{-1,-2,-3\}$ pour les trois phases de lecture du fragment codant complémentaire
- $\mathbb{P}_\pi(COD_0 \mid X_1^w = x_1^w)$ pour le modèle non-codant

Par exemple, on a :

$$\mathbb{P}_\pi(COD_1 \mid X_1^w = x_1^w) = \frac{\mathbb{P}_\pi(X_1^w = x_1^w \mid COD_1) * \mathbb{P}_\pi(COD_1)}{\sum_{f\in\{-1,-2,-3,0,1,2,3\}} \mathbb{P}_\pi(X_1^w = x_1^w \mid COD_f) * \mathbb{P}_\pi(COD_f)}.$$

Une fenêtre de taille w est positionnée en début de séquence. On calcule les sept probabilités *a posteriori* du fragment associées à la position du milieu de la fenêtre. On décale ensuite la fenêtre d'un certain pas s, et on itère le processus. Les valeurs par défaut du programme *GeneMark* sont

[13] `http://bioweb.pasteur.fr/docs/man/man/mkmat.1.html`

$s = 12$ pb et $w = 96$ pb. La sortie graphique est constituée de six panneaux (trois panneaux supérieurs pour le brin direct et trois inférieurs pour le brin inverse). Chaque panneau correspond à une des six manières de lire les triplets consécutifs dans une séquence d'ADN. Les axes verticaux représentent les valeurs de probabilité de codage tandis que les axes horizontaux représentent les positions des nucléotides le long de la séquence d'ADN. Le potentiel de codage d'une CDS en phase f (ou probabilité moyenne de codage (Pc)) correspond à la somme des probabilités[14] $\mathbb{P}_\pi(COD_f \mid X_1^w = x_1^w)$ calculées entre les positions de début et de fin de la CDS, divisée par sa longueur. Ainsi *GeneMark* permet, à partir d'un critère Bayesien, de prédire le modèle (non-codant ou codant) qui s'ajuste le mieux sur une fenêtre glissante de taille fixée. La sélection finale des CDS consiste simplement à utiliser un seuil de probabilité moyenne de codage de 0,5.

Il est important de modéliser l'hétérogénéité de composition en oligonucléotides des CDS révélées par l'analyse de l'usage des codons synonymes (voir p. 197 et p. 98). En théorie, cette modélisation (qui n'est pas réalisée dans *GeneMark*) ne nécessite pas forcément de mettre en œuvre les modèles HMM qui sont à la fois complexes et gourmands en ressources informatiques. En effet, nous avons vu, dans le cas du modèle ombre du codant, qu'il est facile d'étendre la formule de Bayes pour prendre en compte des situations supplémentaires. Il est cependant important de souligner que plus on augmente le nombre de modèles pris en compte par la formule de Bayes, plus la préparation des jeux de séquences pour la phase d'apprentissage est laborieuse. En effet, cette préparation nécessite généralement une expertise manuelle pour chaque modèle de l'espèce procaryote étudiée. Il s'agit d'établir des jeux fiables de CDS homogènes et de séquences non-codantes, c'est-à-dire de trouver le juste équilibre entre la taille de l'échantillon et la confiance que l'on a dans les séquences, autrement dit entre les nombres de faux-négatifs et de faux-positifs.

Programme *Prokov* Le programme *Prokov* ([Romanet, 2001] Tab. 3.3 p. 91) est une implémentation plus étoffée de *GeneMark* et libre d'accès. La méthode *Prokov*, développée par A. Viari *et coll.* repose sur des modèles de chaînes de Markov : un modèle de chaîne de Markov à transitions 3-périodiques (Mm_3) pour le codant, un modèle 3-périodique (Mm_3) pour l'ombre du codant et un modèle homogène (Mm) pour le non-codant. *Prokov* est constitué de cinq modules indépendants pouvant être combinés pour réaliser la phase d'apprentissage et la phase de reconnaissance de gènes procaryotes (Fig. 3.2 p. 97) :

1. *prokov_orf* recherche les CDS maximales d'une certaine longueur dans les six phases de lecture d'une séquence (recherche par signal).
2. *prokov_learn* correspond à la phase d'apprentissage d'une recherche par contenu qui modélise l'ADN codant par des chaînes de Markov. Il génère des matrices d'effectifs de $m + 1$-uple à partir de jeux d'apprentissage par comptage de mots. Le jeu d'apprentissage pour le modèle codant est obligatoire (jeu de CDS issu de *prokov_orf*) alors que le jeu d'apprentissage pour le modèle non-codant est optionnel. Dans le cas où il n'y a pas de jeu d'apprentissage pour le

[14] Le nombre de valeurs de probabilité dépend de la valeur du pas s (idéalement $s = 1$).

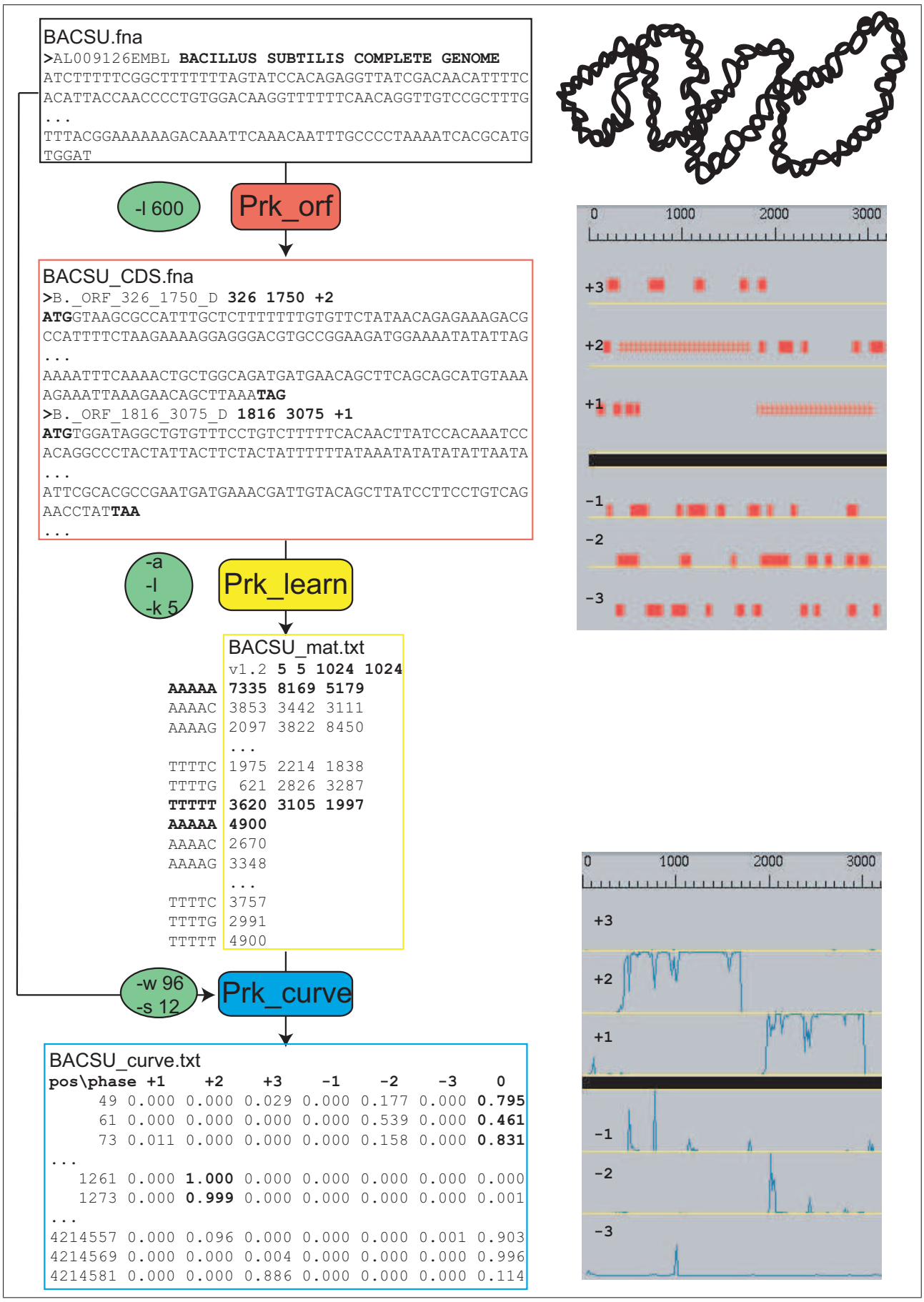


FIG. 3.2 – Les modules Prokov pour la prédiction de gènes procaryotes (annexe C p. 389)

modèle non-codant, *prokov_learn* simule les effectifs du modèle non-codant en mélangeant les effectifs des phases codantes (*shuffle*). Pour fournir à *prokov_learn* un jeu de séquences non-codantes *natives*, il est nécessaire d'extraire du génome étudié les régions intergéniques et/ou les gènes d'ARN fonctionnels, l'idéal étant d'avoir deux modèles d'ADN non-codant plutôt qu'un seul qui regroupe les régions intergéniques et les gènes d'ARN fonctionnels. Dorénavant, pour plus de simplicité, nous regrouperons sous le terme unique de *matrice de probabilités de transition* l'ensemble des quatre matrices : celles dérivées pour les trois positions des codons à partir d'un jeu de CDS (celles décrivant le modèle ombre du codant se déduisent de ces dernières), et celle décrivant le modèle non-codant (générée par *shuffle* ou à partir d'un jeu de séquences non-codantes réelles).

3. *prokov_curve* applique la formule de Bayes sur une fenêtre glissante pour calculer les probabilités de codage le long de la séquence en utilisant la matrice de probabilités de transition préalablement générée par *prokov_learn* (phase de reconnaissance de la recherche par contenu). *prokov_learn* calcule, en une seule lecture, les probabilités *a posteriori* des fragments contenus dans la fenêtre : $\mathbb{P}_\pi(COD_m \mid X_1^w = x_1^w),\ f \in \{-1,-2,-3,0,1,2,3\}$.
4. *prokov_score* applique la formule de Bayes directement sur un jeu de CDS fourni par l'utilisateur (pouvant provenir de *prokov_orf*) pour calculer leur score de codage en utilisant une matrice de transitions ($\mathbb{P}_\pi(COD_f \mid X_1^L = x_1^L),\ f \in \{-1,-2,-3,0,1,2,3\}$, où L est la longueur de la CDS. Phase de reconnaissance alternative à *prokov_curve*).
5. Enfin, *prokov_cds* permet de combiner automatiquement les modules *prokov_orf* et *prokov_score* en utilisant une matrice de transitions.

Programme *GeneMark-Genesis* Le problème des méthodes fondées sur un modèle d'ADN codant est la phase d'apprentissage : il est nécessaire de fournir au programme au minimum un jeu de CDS, voire un jeu de séquences non-codantes. Par ailleurs, les travaux de C. Médigue [Médigue *et al.*, 1991] ont révélé l'existence d'une troisième classe de gènes dans le génome de *E. coli* K-12, participant à l'hétérogénéité compositionnelle en oligonucléotides des CDS : l'analyse des fonctions biologiques de ces gènes suggère fortement qu'ils sont impliqués dans des mécanismes de transferts horizontaux. Les travaux de M. Borodovsky ont montré que la modélisation des différentes classes de gènes (gènes typiques, gènes hautement exprimés et gènes impliqués dans les transferts horizontaux) permettait d'améliorer la sensibilité de la prédiction de gènes (moins de faux-négatifs [Borodovsky *et al.*, 1995]). En d'autres termes, il est possible d'itérer les programmes du type *GeneMark* ou *Prokov* k fois, k correspondant au nombre de classes de gènes décrites chez un organisme. A chaque itération, on fournit au programme un modèle de gènes homogènes dans leur usage des codons synonymes. Ce raffinement des étapes d'apprentissage et de reconnaissance nécessite le développement d'un programme de construction automatique de modèles générant à la fois un jeu de CDS et un jeu de séquences non-codantes, et proposant plusieurs ensembles de CDS associés aux différentes classes de gènes homogènes. C'est dans le but de répondre à ces attentes

que le programme *GeneMark-Genesis* a été développé (Tab. 3.3 p. 91). La méthode est constituée de deux phases principales :

1. construction d'une matrice de transitions de gènes et de séquences non-codantes
2. construction d'une matrice de transitions pour chaque groupe de gènes homogènes dans leur usage des codons synonymes

La première phase comporte deux étapes :

1. construction d'une prématrice classique
2. construction d'une matrice classique

A partir de la séquence du chromosome, *GeneMark-Genesis* extrait les plus longues CDS d'une taille supérieure à 700 pb ; ce jeu de CDS permet de construire des prématrices de transition en utilisant un modèle $M0$ pour le non-codant (fréquences en nucléotides du génome). Les prématrices sont utilisées par le programme *GeneMark* pour prédire les CDS du chromosome. Du jeu de CDS prédites est déduit le jeu de séquences non-codantes par soustraction des CDS à la séquence du chromosome complet. Ces jeux de CDS et de séquences non-codantes permettent de construire une matrice de transition pour le modèle d'ADN codant et non-codant.

La seconde phase comporte trois étapes :

1. partitionnement automatique de CDS en k groupes homogènes
2. construction d'une prématrice pour chaque groupe de gènes
3. construction d'une matrice pour chaque groupe de gènes

La méthode de partitionnement automatique utilisée est une variante du K-means (voir p. 138). Nous prendrons l'exemple du 2-means qui sépare les CDS en deux groupes (typique et atypique), plutôt que l'exemple du 3-means (typique, atypique et hautement typique) puisque W. Hayes conclut que la division en deux groupes est amplement suffisante dans le cadre de la prédiction de gènes [Hayes & Borodovsky, 1998b].

1. L'initialisation consiste à définir deux groupes de CDS, de la façon suivante. La matrice issue de la première phase est utilisée par *GeneMark* pour calculer la probabilité moyenne de codage des CDS de longueur supérieure à 700 pb. Celles dont la probabilité moyenne de codage est supérieure ou égale à 0,5 constituent le groupe initial de CDS typiques. Le reste constitue le groupe atypique. Puis on calcule l'usage relatif des codons synonymes (*Relative Synonymous Codon Usage* (*RSCU*)) des deux tableaux multidimensionnels (typique et atypique) à n CDS et 59 codons. Pour cela, la fréquence du codon *abc* est normalisée par celle de l'acide aminé encodé et multipliée par la taille du groupe de codons synonymes (Tab. 3.2 p. 89).
2. Les itérations se déroulent en deux temps. Premièrement, le centre de chacun des groupes est défini par le vecteur des 59 RSCU cumulatives, observées dans toutes les CDS du groupe. Deuxièmement, on réaffecte toutes les CDS au centre le plus proche selon une fonction de distance de type Kullback-Liebler [Hayes & Borodovsky, 1998b, Besemer & Borodovsky, 1999].

La formule Kullback-Liebler permet de mesurer le *contraste* entre deux modèles statistiques en comparant leur contenu en information. Elle est fondée sur la somme des logarithmes des ratios des vraisemblances contenues dans les deux modèles (*log-likelihood ratio*). Elle est encore appelée entropie relative. Cet indice de distance n'est pas une distance euclidienne puisqu'elle ne respecte pas l'inégalité triangulaire.

3. La convergence est atteinte lorsque les groupes sont stables

Ainsi on obtient deux groupes de CDS homogènes qui sont utilisés, avec le modèle non-codant $M0$, pour construire deux prématrices (typique et atypique).

Les CDS prédites par au moins l'une de ces deux matrices et présentes dans le jeu de séquences non-codantes issu de la première phase sont éliminées de ce dernier (elles correspondent à des faux-négatifs). La matrice finale typique (respectivement atypique) est dérivée à partir du groupe de CDS typiques (respectivement atypiques), et du jeu nettoyé de séquences non-codantes.

W. Hayes souligne qu'utiliser des CDS de taille supérieure à 700 pb pour la phase d'apprentissage n'affecte pas la précision de prédiction des petites CDS pouvant présenter d'éventuels biais dans leur usage des codons synonymes relativement aux longues CDS. Cette méthode donne de bons résultats, mais nécessite de fixer un certain nombre de paramètres au départ (le nombre de groupes) et les étapes sont relativement complexes. Finalement *GeneMark-Genesis*[15] permet de construire un modèle typique et un modèle atypique à partir d'une séquence chromosomique (phase d'apprentissage), ensuite utilisés par *GeneMark* ou *GeneMark.hmm* (voir p. 114) pour prédire les CDS sur cette même séquence (phase de reconnaissance).

Programme développé par Audic et Claverie Cette méthode qui n'a pas de nom repose sur les modèles de chaînes de Markov d'ordre m ([Audic & Claverie, 1998] Tab. 3.3 p. 91). La méthode comporte deux phases principales :

1. modèle de chaînes de Markov homogènes
2. raffinement du modèle par chaînes de Markov 3-périodiques

Chacune de ces phases permet un apprentissage automatique et une prédiction de gènes procaryotes par un processus itératif en deux étapes :

1. construction d'une matrice de transitions
2. partitionnement des séquences génomique en trois groupes : codant sur le brin direct, codant sur le brin inverse (ombre) et non-codant

La première phase se déroule en trois étapes :

1. Lors de l'initialisation, la séquence génomique est découpée aléatoirement en segments de longueur $w = 100$ pb non chevauchants. Ces segments sont répartis aléatoirement en trois groupes G_1, G_2, G_3.

[15] `opal.biology.gatech.edu/GeneMark/downloads.html`

2. Les itérations se passent en deux temps. Premièrement, on construit une prématrice de transition à partir des trois groupes de segments ($M5$). Deuxièmement, on applique la formule de Bayes sur une fenêtre glissante ($w = 100$ et $s = 5$) pour calculer les trois probabilités *a posteriori* $\mathbb{P}_\pi(G_m \mid X_1^w = f_1^w),\ m \in \{1,2,3\}$ le long de la séquence en utilisant la prématrice de transitions. On définit des segments de séquence et on les réaffecte en trois groupes. Chaque fenêtre est tout d'abord affectée au groupe sur lequel elle obtient la meilleure probabilité $\mathbb{P}_\pi(G_m \mid X_1^w = f_1^w)$. Si au moins $w/s = 20$ fenêtres consécutives appartiennent au même groupe alors on affecte le segment démarrant au milieu de la première des fenêtres et se terminant au milieu de la dernière fenêtre au groupe m (sinon, le segment n'est pas classé).
3. La convergence est atteinte quand les groupes ne varient presque plus, c'est-à-dire quand la taille et le contenu de chaque groupe ne varient pas significativement (<0,5%) d'une itération à la suivante.

Les auteurs précisent que cette convergence s'obtient en général en moins de 50 itérations et que la répartition aléatoire de l'initialisation n'influe pas sur la convergence. A ce stade, on dispose donc de trois groupes *anonymes* c'est-à-dire que l'on ne sait pas lequel correspond à des séquences codantes sur le brin direct et sur le brin complémentaire et aux séquences non-codantes. Dans la seconde phase on recommence le même type de procédure que dans la première phase :

1. Lors de l'initialisation, une fonction est attribuée à chacun des trois groupes de la première phase de manière *ab initio* (ou instrinsèque). On calcule, pour chaque groupe, la proportion de séquences contenant des ORF (sans codon de terminaison en phase) sur chacune des phases $-1, -2, -3, 1, 2, 3$. Le groupe obtenant la plus grande proportion sur les phases positives correspond au codant sur le brin direct ; réciproquement, le groupe obtenant la proportion la plus grande sur les phases négatives correspond au codant sur le brin complémentaire, et le groupe restant correspond au non-codant.
2. Les itérations se passent en deux temps. Premièrement, on extrait les ORF du groupe de segments codants directs et du groupe de segments codants inverses. On obtient ainsi trois groupes : les ORF codantes directes, les ORF codantes inverses et les segments non-codants. On construit une matrice de transition à partir de ces trois groupes ($M5\_3$ pour les deux groupes de codant et $M5$ pour le non-codant). Deuxièmement, on applique la formule de Bayes sur une fenêtre glissante pour calculer les sept probabilités $\mathbb{P}_\pi(COD_f \mid X_1^w = x_1^w),\ f \in \{-1, -2, -3, 0, 1, 2, 3\}$ le long de la séquence, en utilisant la matrice de transition. On définit des segments de séquence et on les réaffecte en trois groupes de la même la manière que pour la première phase.
3. La convergence est atteinte quand les groupes ne varient presque plus.

Cette méthode a l'avantage de n'utiliser qu'un minimum de connaissances *a priori* sur les modèles étudiés (k et w). Le modèle de chaînes de Markov de la seconde phase est le même que celui de *GeneMark*. A la différence de *GeneMark-Genesis*, il n'est pas possible ici de construire plusieurs matrices de transitions homogènes dans l'usage des codons synonymes des CDS, bien

qu'une modification du programme soit envisageable dans ce but. Par ailleurs, le processus itératif (dans la première et la seconde phase) peut être considéré comme une technique de partitionnement automatique des séquences génomiques en trois classes qui s'inspire de l'algorithme EM (voir p. 105). Nous mentionnons à cet effet la méthode EMKOV qui utilise explicitement la technique EM [Romanet, 2001]. EMKOV emprunte des principes à la fois à la méthode *GeneMark-Genesis* et à la méthode développée par Audic et Claverie puisqu'elle vise

1. à ne pas utiliser d'ensemble d'apprentissage donné *a priori*
2. à produire les matrices par raffinements successifs
3. à produire potentiellement plusieurs ensembles de matrices associées à des catégories de gènes différentes

Elle est, par ailleurs, beaucoup plus simple et fait intervenir moins de paramètres (elle est non supervisée puisque l'utilisateur n'a pas à définir le nombre de classes) et moins d'étapes que ces deux méthodes.

Programmes *GLIMMER* et *FrameD* L'estimation des paramètres d'un modèle markovien d'ordre m nécessite de compter les $m+1$-uplets observés dans un jeu d'apprentissage suffisamment grand. Plus l'ordre du modèle est élevé, meilleure est la précision de la reconnaissance de gènes sur une nouvelle séquence d'ADN, mais plus le jeu d'apprentissage doit être grand pour estimer correctement les comptages du $m+1$-uple. Par ailleurs, certains $m+1$-uplets pouvant être sous-représentés et d'autres sur-représentés, l'utilisation d'un ordre unique n'est pas nécessairement le meilleur choix. D'autres méthodes de prédiction de gènes sont capables d'utiliser des modèles à dépendance variable désignés sous le terme général d'arbres de contexte. Par exemple, les chaînes de Markov à longueur variable (ou ordre variable) sont un cas particulier d'arbres de contexte [Nuel, 2001]. Elles sont définies comme une généralisation des chaînes de Markov d'ordre fixe qui combine des contextes de différentes longueurs pour calculer la probabilité d'apparition d'une base.

Le programme *GLIMMER* 1.0, développé au TIGR par S. Salzberg [Salzberg *et al.*, 1998], utilise des *Interpolated Markov Model* (*IMM*) d'ordre 0 à 8. Son auteur décrit un modèle markovien 3-périodique pour l'ADN codant du brin direct, un modèle 3-périodique pour le codant du brin inverse et un modèle homogène pour le non-codant. On calcule la probabilité d'apparition d'une CDS par $\mathbb{P}_\pi(X_1^L = x_1^L \mid COD_f) = \sum_{i=1}^{L} IMM_8(X_i),\ f \in \{-1,-2,-3,0,1,2,3\}$, où X_i est l'oligomère de longueur 9 $(k+1)$ se finissant à la position i. Cette probabilité conditionnelle est donc la somme des scores IMM d'ordre 8 de tous les oligomères. Le score IMM_8 de l'oligomère X_i est calculé par $IMM_m(X_i) = \lambda_m(X_{i-1}) * P_m(X_i) + [1 - \lambda_m(X_{i-1})] * IMM_{m-1}(X_i)$, où $\lambda_m(X_{i-1})$ est le poids[16] associé au m-uplet finissant en position $i-1$ de la séquence X, et $P_m(X_i)$ la probabilité de la base en position i de la séquence dans un modèle d'ordre m. Les probabilités de transition sont estimées par maximum de vraisemblance pour les $m+1$-uplets (m variant de 0 à 8), à partir d'un jeu d'apprentissage constitué de séquences codantes. Ainsi, le score IMM_8 d'un oligomère

[16] Le terme interpoler signifie que le score est pondéré par un coefficient ici λ.

est la combinaison linéaire des prédictions faites par les modèles d'ordre 8, 7, 6, etc, jusqu'au modèle d'ordre 0. Si l'effectif du m-uplet est suffisant, le poids $\lambda_m(X_i - 1)$ est égal à 1, sinon les fréquences observées d'ordre m sont comparées[17], à l'aide d'un χ^2, aux probabilités IMM d'ordre $m-1$ précédemment calculées. Si les valeurs diffèrent de manière significative, le poids est égal à $\sum_{a\in\{A,C,G,T\}} \pi_m(X_{i-1}, a)$, sinon il est nul (les fréquences observées n'offrent qu'une faible valeur prédictive). Chaque contexte va ainsi être en partie pondéré par sa fréquence, ce qui permet de choisir par exemple un ordre 8 pour les 9-uplets suffisamment fréquents, et un ordre 5 ou inférieur pour les 9-uplets rares. Autrement dit, les IMM utilisent les prédictions de modèles d'ordre inférieur pour lesquels les données sont plus importantes, afin d'ajuster les prédictions faites à partir de modèles d'ordre supérieur.

Le programme *GLIMMER* 2.0 ([Delcher *et al.*, 1999a] Tab. 3.3 p. 91) utilise des arbres de contexte interpolé (*Interpolated Context Model* (*ICM*) ; contexte à trou extension des IMM [Nicolas, 2003]) d'ordre 0 à 12 (ICM_{12}). Pour un contexte donné $x_1x_2\ldots x_m$ de longueur m, les IMM de *GLIMMER* 1.0 calculent la probabilité de x_{m+1} en utilisant autant de nucléotides, précédant immédiatement x_{m+1}, que le jeu d'apprentissage le permet. Les ICM sont plus flexibles car ils permettent de sélectionner *n'importe quels* nucléotides du contexte pour déterminer la probabilité de x_{m+1} (et pas seulement ceux qui sont adjacents à x_{m+1}).

Le programme *GLIMMER* se présente, comme *Prokov*, sous la forme de plusieurs modules (*long-orf*, *extract*, *build-imm*, *glimmer2*) combinés dans le script *run-glimmer2* qui permet d'automatiser la phase d'apprentissage et la phase de reconnaissance d'une séquence d'ADN suffisamment longue (chromosome procaryote).

La phase d'apprentissage de *GLIMMER* 1.0 et 2.0 nécessite uniquement un jeu de CDS (le modèle de probabilités indépendantes $M0$ est utilisé pour le non-codant). Dans le script *run-glimmer2*, le jeu de CDS pour l'apprentissage (module *build-imm*) est simplement constitué des CDS maximales, non chevauchantes et de longueur supérieure à 500 pb, extraites du chromosome avec les modules *long-orf* et *extract*.

On dit que la phase de reconnaissance se fait par *assimilation* puisque *GLIMMER* ne connaît que les caractéristiques du codant. Cette seconde phase effectuée par le module *glimmer2*, comporte deux étapes :

1. le calcul du potentiel de codage de toutes les CDS du génome de longueur supérieure ou égale à 90 pb
2. la sélection des CDS selon plusieurs critères

Le score d'une CDS mesure la vraisemblance de la CDS sous le modèle codant en phase 1, normalisée par les vraisemblances de la CDS dans les autres modèles[18] ($\mathbb{P}_\pi(COD_1 \mid X_1^L = x_1^L)$;

[17] Plus précisément, $\pi_m(X_{i-1}, A)$, $\pi_m(X_{i-1}, C)$, $\pi_m(X_{i-1}, G)$ et $\pi_m(X_{i-1}, T)$ sont comparées, respectivement à, $IMM_{m-1}(X_i, A)$, $IMM_{m-1}(X_i, C)$, $IMM_{m-1}(X_i, G)$ et $IMM_{m-1}(X_i, T)$.

[18] Le score est normalisé par les sept vraisemblances dans le cas d'une CDS courte (inférieure à 500 pb). Le modèle non-codant n'est pas utilisé dans la normalisation d'une CDS longue. On considère implicitement que toute CDS longue est codante.

formule de Bayes ; [Schiex *et al.*, 2000]). S. Salzberg considère qu'utiliser simplement un seuil de codage pour la sélection des CDS n'est pas une méthode satisfaisante [Salzberg *et al.*, 1998].

Il est nécessaire, d'une part, de définir un seuil de codage en dessous duquel aucune région ne peut être considérée comme codante, et d'autre part, de tenir compte du problème difficile de la gestion des gènes recouvrants. *GLIMMER* 2.0 résout les conflits de recouvrement entre deux gènes A et B en choisissant progressivement un codon d'initiation plus en $3'$ pour le gène A et/ou pour le gène B en fonction des cas de figure (indépendamment de la présence d'un RBS).

Généralement, pour réajuster le codon d'initiation des CDS en fonction des RBS, on utilise le programme *RBSfinder* dans une étape de post-traitement, mais on peut aussi choisir, dès le départ, le codon d'initiation en fonction du RBS au lieu de prendre celui qui se trouve le plus en $5'$ (option $-f$[19] de *glimmer2*). La méthode *GLIMMER* permet d'améliorer la reconnaissance de zones codantes en utilisant une combinaison de modèles à dépendance variable. Cependant, elle ne tient pas compte de l'hétérogénéité de la composition en oligonucléotides des gènes (les gènes atypiques risquent d'être négligés). Dans *GLIMMER* 2.0, A. Delcher résout partiellement ce problème en réalisant un deuxième passage sur la séquence, destiné à *boucher les trous* d'annotations. La description de la liste des gènes prédits contient un statut pour les gènes dont l'annotation est délicate : *doubtful overlap* pour étiqueter des gènes recouvrants, *prospective gene* pour les gènes identifiés au deuxième passage, etc. Le programme *GLIMMER* est plus sensible que *GeneMark* (il génère moins de faux-négatifs) mais il est, en contrepartie, moins spécifique (il génère plus de faux-positifs). La nouvelle version *GLIMMER* 2.10 modélise le non-codant, ce qui conduit à une réduction du taux de faux-positifs.

Les contraintes de codage et de richesse en G+C entraînent conjointement la création de longues ORF qui sont des *miroirs* des véritables gènes [Schiex *et al.*, 2000]. Naturellement, de telles ORF ont généralement un caractère non-codant. *GLIMMER* par exemple, s'appuie de façon cruciale sur cette notion de phase de lecture ouverte de grande taille, et produit des résultats de mauvaise qualité sur les génomes riches en G+C (*M. tuberculosis* H37Rv). T. Schiex a donc été amené à construire le programme de prédiction de gènes *FrameD* ([Schiex *et al.*, 2003] Tab. 3.3 p. 91), qui reprend les ingrédients de base de *GLIMMER*, mais qui s'appuie plus fortement sur les IMM pour décider si une ORF est ou n'est pas un gène. Il utilise un graphe dirigé sans circuit (*Directed Acyclic Graph* (*DAG*)) qui modélise les CDS, les RBS, les décalages du cadre de lecture et les recouvrements entre deux gènes dans le même sens. Chaque nucléotide est représenté par sept arrêtes pondérées par les sept probabilités d'émission d'un nucléotide, et par un score de similarité si le programme Blastx a permis de mettre en évidence une similitude avec les protéines répertoriées dans les banques de séquences. L'utilisateur doit fournir un jeu de CDS pour la phase d'apprentissage. La phase de reconnaissance de gènes permet d'ajuster le codon d'initiation en fonction de la force du RBS en $5'$. La recherche du plus court chemin dans le graphe utilise un algorithme de programmation dynamique (algorithme de Bellman [Schiex *et al.*, 2000]).

[19] Cette option n'est pas complètement testée. La reconnaissance du RBS est une recherche exacte de chaîne de caractères.

Méthodes fondée sur les modèles de chaînes de Markov cachées

Les programmes de prédiction de gènes fondés sur les modèles de chaînes Markov simples ou interpolées nécessitent une phase d'apprentissage et une phase de reconnaissance. La phase de reconnaissance de *GeneMark* effectue :

1. une mesure des probabilités de codage le long de la séquence au moyen d'un fenêtre glissante
2. une recherche de CDS par signal
3. un calcul de probabilité moyenne de codage associée à chaque CDS
4. une sélection des CDS selon leur probabilité moyenne de codage

La phase de reconnaissance de *GLIMMER* effectue

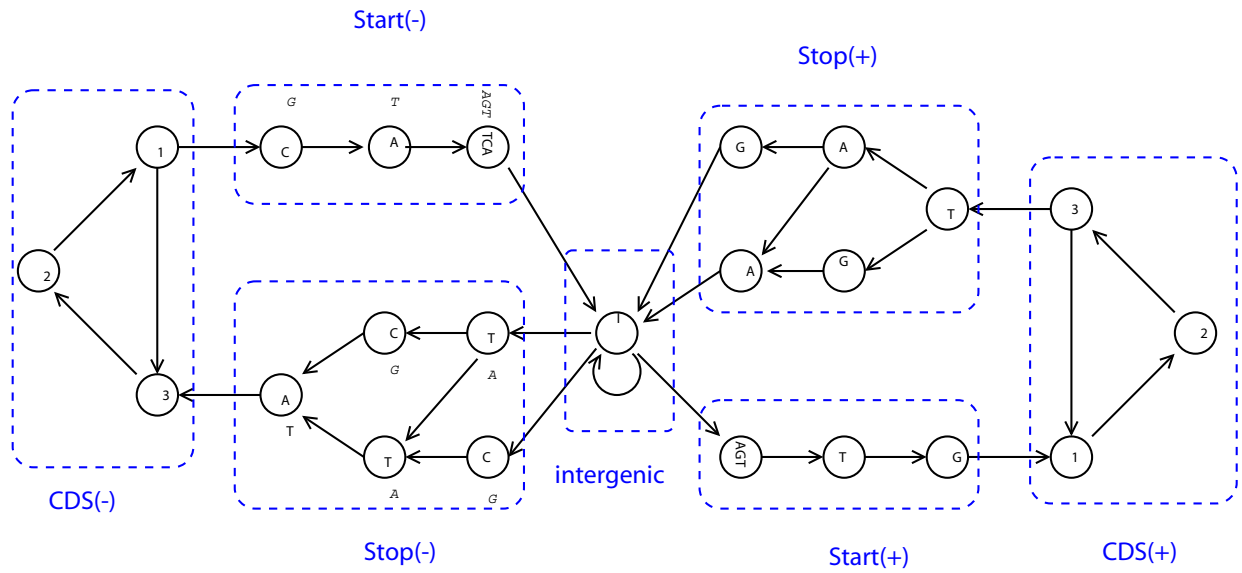
1. une recherche de CDS par signal
2. une mesure du potentiel de codage directement sur les CDS
3. une sélection des CDS selon plusieurs critères

Aucune de ces deux approches ne permet de localiser simplement les frontières entre régions codantes et non-codantes (elles nécessitent des étapes de post-traitements pour sélectionner les CDS). D'un point de vue probabiliste, la façon naturelle de résoudre ce problème, consiste à modéliser non seulement la composition en oligonucléotides des régions codantes et non-codantes, mais aussi l'alternance de ces régions. Cette modélisation permet alors de répondre à la question : « où sont les régions de différents types (intergénique, CDS directe, CDS inverse) ? » et non plus : « étant donnée une région, de quel type est-elle ? ».

Principe des HMM Les modèles de chaînes de Markov cachées (*Hidden Markov Models* (*HMM*), généralisation des chaînes de Markov ; [Rabiner, 1989]), ont été introduits pour la première fois dans les problèmes de reconnaissance de la parole. Appliqués à la prédiction de gènes, ils permettent de modéliser la séquence d'ADN comme une mosaïque de régions homogènes d'un nombre de types relativement restreint, possédant chacunes leures propriétés de composition. Le modèle le plus simple permettant de représenter l'alternance des régions est, une fois de plus, le modèle de chaînes de Markov. Dans ce modèle, le type de région à chaque position de la séquence est généré en ne tenant compte que du type de région à la position précédente ($M1$). Cette chaîne est dite cachée puisqu'elle n'est observable qu'à travers les hétérogénéités de la composition de la séquence d'ADN. Les types de régions sont appelés états (ou régimes) cachés du modèle. L'apparition ou l'émission des nucléotides le long de la séquence est modélisée conditionnellement à la suite des états cachés. Un HMM peut être une succession aléatoire d'un petit nombre de chaînes de Markov (leur succession étant également markovienne). Les HMM constituent donc un cadre simple et souple de modélisation de l'alternance de différents types de régions homogènes au sein de la séquence d'ADN.

Un HMM se caractérise par l'emboîtement de deux processus :

1. celui, observable, de la séquence $X_1^n = X_1 \dots X_n$ où $X_i \in \mathcal{A}$



This figure displays the structure of a simple HMM (represented by the hidden state transitions) for which SHOW has been designed for. Circles represent the considered states of the HMM and arrows the allowed transitions between states. This graph modelises the alternance of intergenic regions with CoDing Sequences (CDS) on the both strands of a DNA molecule. The model contains 23 hidden states grouped together in 7 groups (in dotted line). Here follows a short description of the biological meaning of this graph (we only give details of the model structure).

- When the actual hidden state corresponds to intergenic region at position t, arrows indicate that we can stay in at position $t+1$ or leave it towards the first position of start codon on the (+) strand (triplets: atg, gtg, ttg) or towards a last position of a stop codon on the (−) strand (inverse complementary of tga, tag, taa). Intergenic state can only be reached from the last position of a stop codon on the (+) strand or the first position of a start codon on the (−) strand.

- Leaving the third position of a start codon on the (+) strand, the hidden path go through a CDS which is a succession of codons. Codons are modelized by using a cycle of three hidden states, one for each position inside the codon. This modelisation ensures that the length of the CDS will be a multiple of three and allows to take into account distinct compositions of the DNA according to the codon position (defined by the emission transitions π_o which are not described here).

- At the third codon position on the (+) strand, it is necessary to forbid the appearance of an in frame stop codon. Thus, the emission model associated to the third position needs to verify: $\pi_o(\text{CDS+\_3}, tg, a) = P(X_i = a \mid S_i = \text{CDS+\_3}, X_{i-2}^{i-1} = tg) = 0$, $\pi_o(\text{CDS+\_3}, ta, g) = P(X_i = g \mid S_i = \text{CDS+\_3}, X_{i-2}^{i-1} = ta) = 0$ and $\pi_o(\text{CDS+\_3}, ta, a) = P(X_i = a \mid S_i = \text{CDS+\_3}, X_{i-2}^{i-1} = ta) = 0$.

- The first position of the stop codon on the (+) strand could only be reached from the third position of a translated codon and corresponds to the first nucleotide of the triplets: tga, tag, taa.

- From the third stop codon position on the (+) strand, the path goes through intergenic.

- Being in intergenic state, in a symmetric way as for (+) strand CDS, the third position of a stop codon on the (−) strand is the beginning of a (−) strand CDS.

FIG. 3.3 – Exemple d'un HMM simple dédié à la prédiction de gènes bactériens
Cette figure a été extraite du manuel utilisateur du programme *SHOW* [Nicolas, 2003].

2. celui des états cachés $S_1^n = S_1 \dots S_n$ qui gère l'alternance des plages homogènes, chacun des S_i prenant sa valeur dans $\mathcal{S} = \{1, \dots, q\}$

Dans une séquence $(X_i, S_i)_{0 \leq i \leq n}$, les S_i réalisent une segmentation de cette dernière : à l'intérieur d'un segment toutes les lettres X_i sont tirées selon la même loi markovienne. La reconstruction des régimes à partir de la séquence observée permet d'identifier q régions homogènes distinctes dans la séquence, chaque région pouvant se répartir sur différentes plages de la séquence et ainsi révéler son hétérogénéité. On peut choisir, par exemple, de décrire les séquences d'ADN par trois modèles : celui des gènes du brin direct, celui des gènes du brin reverse et les régions intergéniques. Le modèle complet, présenté dans la figure 3.3 p. 106, boucle sur lui-même de manière à pouvoir détecter l'ensemble des gènes sur les deux brins : on passe du modèle intergénique au modèle de gène direct ou inverse, pour revenir au modèle intergénique, etc. Chaque modèle peut comporter plusieurs sous modèles constitués de différents états cachés. Par exemple, le modèle de CDS du brin direct peut être décrit par trois sous-modèles : codon d'initiation direct, codons sens du brin direct et codon de terminaison direct. Le sous modèle des codons sens du brin direct décrit trois états cachés dépendant de la position des nucléotides dans le codon. Lorsque l'on arrive à l'état caché du troisième nucléotide d'un codon sens, on a le choix entre boucler (revenir) sur l'état caché du premier nucléotide (et ainsi allonger la chaîne de codons sens), ou bien passer à l'état du premier nucléotide du sous modèle du codon de terminaison. Le modèle complet est donc constitué de sept sous modèles (vingt-trois états cachés) : intergénique (un état qui boucle ; Mm), codon d'initiation direct (trois états ; $M0$, fréquence relative des bases), codons sens sur le brin direct (trois états qui bouclent ; Mm), codon de terminaison direct (cinq états[20] ; $M0$), et symétriquement sur le brin inverse, codon d'initiation, codons sens, codon de terminaison. Les paramètres de ce modèle (θ) sont regroupés dans

1. une matrice de probabilités de transition entre états cachés ($M1$)
2. vingt-trois matrices de probabilités d'émission des nucléotides au sein des états cachés ($M0$ et Mm)

De nombreuses probabilités de transition entre états cachés peuvent être fixées à 0, permettant ainsi de structurer la suite des états cachés (graphe des transitions autorisées entre états cachés ; Fig. 3.3 p. 106). Quel que soit le modèle HMM, deux phases principales sont nécessaires à la prédiction de régions homogènes d'une séquence :

1. La phase d'apprentissage estime tous les paramètres du modèle $\theta = (\Pi_S, \Pi_X)$ afin de caractériser chacun des états (à partir de leur composition statistique en oligonucléotides) et de définir un modèle de séquence aléatoire *ressemblant* à la séquence observée. Dans le cas de l'observation de données complètes $Y_1^n = (X_1^n, S_1^n)$, la situation est simple : il est possible de calculer explicitement l'estimateur du maximum de vraisemblance de l'ensemble des

[20] On remarque que la modélisation des trois codons de terminaison TAA, TAG, TGA est plus compliquée que celle des trois codons d'initiation [AGT]TG. Dans le cas où l'on a un état caché par nucléotide, les codons de terminaison sont modélisés par cinq ou six états cachés.

paramètres. Généralement on observe uniquement la séquence biologique, et non les états cachés. L'estimateur du maximum de vraisemblance n'est alors pas calculable explicitement car la combinatoire générée par tous les découpages possibles de la séquence est beaucoup trop importante. L'algorithme *Expectation Maximisation* (*EM* [Rabiner, 1989]) est un algorithme général d'estimation de paramètres par maximum de vraisemblance (*Maximum Likelihood* (*ML*)) dans le cas de modèles à données incomplètes. Autrement dit, il permet d'approcher l'estimateur du ML lorsque la maximisation directe du log-vraisemblance n'est pas possible.

2. La phase de reconnaissance segmente une nouvelle séquence d'après les paramètres estimés afin de restaurer la suite des états cachés. Le terme de *décodage* est employé dans le jargon de la reconnaissance de la parole pour désigner la structure sous-jacente de la suite d'états cachés qui donne un sens à la séquence observée [Durbin *et al.*, 2001d]. Par analogie, on peut utiliser le terme de *décodage* pour désigner le processus qui trouve l'analyse optimale de la séquence d'ADN en région codantes et non-codantes [Larsen & Krogh, 2003]. Il existe différentes approches de décodage. L'algorithme de Baum-Welch, qui est l'algorithme EM[21] appliqué aux chaînes de Markov cachées, permet de calculer les probabilités *a posteriori* des q différents états cachés à chaque position i de la séquence ; c'est-à-dire $\mathbb{P}_\theta(S_i = u \mid X_1^n = x_1^n)$, la probabilité que l'observation x_i vienne de l'état u connaissant la séquence observée et les paramètres du modèle (il ne reste plus qu'à définir un seuil de prédiction [Rabiner, 1989, Durbin *et al.*, 2001d]). L'algorithme de Viterbi trouve le découpage le plus probable, c'est-à-dire qu'il reconstruit la suite des états cachés qui correspond le mieux à la séquence observée : $s_1^n* = argmax_S \mathbb{P}_\theta(S_1^n = s_1^n \mid X_1^n = x_1^n)$ avec s_1^n* le chemin le plus probable (cet algorithme ne fournit pas de signification statistique sur chacune des prédictions [Rabiner, 1989, Durbin *et al.*, 2001d]).

Les phases d'apprentissage et de reconnaissance sont, en pratique, regroupées au sein de l'algorithme EM qui permet à la fois de reconstruire les états cachés, et d'estimer les paramètres θ. Cet algorithme déterministe de maximisation itérative de la vraisemblance, fondé sur la programmation dynamique, a pour principe général :

1. Initialisation des paramètres du modèle (valeur arbitraire $\theta^{(0)}$).
2. Itérations en deux temps (itération $j+1$) : segmentation de la séquence pour restaurer la suite des états cachés à partir de la séquence observée et des paramètres courants $\theta^{(j)}$ (étape E ; algorithme *Forward-backward*), et actualisation de $\theta^{(j)}$ en $\theta^{(j+1)}$ en maximisant la vraisemblance de la séquence complète (étape M).
3. Terminaison : la convergence (c'est-à-dire la différence de vraisemblances entre deux itérations est inférieure à un seuil donné) où le nombre d'itération fixé est atteint.

L'utilisation d'un HMM facilite la modélisation de l'hétérogénéité de composition en oligonucléotides des CDS. Théoriquement, l'étape d'apprentissage est complètement automatique et l'étape de reconnaissance ne nécessite ni fenêtre glissante, ni post-traitements (recherche de RBS).

[21] EM est un algorithme plus général que Baum-Welch qui peut être utilisé pour différents modèles probabilistes

Programmes *EcoParse* et *EasyGene* Un premier exemple de programme qui implémente les modèles HMM dans le cadre de la prédiction de gènes procaryotes est *EcoParse* ([Krogh *et al.*, 1994] Tab. 3.3 p. 91). Il a été initialement développé dans le cadre de l'étude du génome d'*E. coli* K-12. A. Krogh modélise, sur un brin de la séquence,

1. les codons d'initiation et de terminaison
2. les régions codantes autorisant des décalages du cadre de lecture
3. des petits chevauchements entre deux gènes dans le même sens (chevauchements sur les codons de terminaison et d'initiation : T[AG]ATG, [AGT]TGA)
4. des patrons de régions intergéniques (séquences palindromiques extragéniques répétées et motifs RBS)

Le processus caché est complexe car il compte de nombreux états cachés (681 au total). En revanche, les processus observés sont décrits par le modèle simple prenant en compte les fréquences relatives des nucléotides $M0$.

La phase d'apprentissage nécessite une étape d'extraction de jeux de séquences codantes et non-codantes (à partir d'un jeu d'annotations du chromosome), et n'est donc ni complètement automatique, ni complètement *ab initio*. Les probabilités d'émission d'un nucléotide au sein des états cachés des modèles intergéniques sont estimées par l'algorithme *Forward-backward* sur le jeu de séquences non-codantes. Les probabilités de transition dans un des 61 modèles de codon sont proportionnelles aux fréquences relatives des codons ($F_R(abc)$) de la table de l'usage du code génétique, établie sur le jeu de CDS.

La phase de reconnaissance utilise l'algorithme de Viterbi pour trouver le chemin le plus probable à travers les états cachés et prédire ainsi les gènes. Pour donner une valeur de confiance à ces prédictions, A. Krogh mesure le potentiel de codage d'une CDS sous le modèle de codons aléatoires et indépendants ($F_R(abc)$). Il définit un indice de codage fondé sur le logarithme de la probabilité d'une CDS sous le modèle codant en phase 1 ($\mathbb{P}_\pi(X_1^l = c_1^l \mid COD_1) = \prod_{j=1}^{l} F_R(a_j b_j c_j)$, avec l la longueur en codons d'une CDS). La phase de reconnaissance se déroule en trois étapes : analyse de la séquence directe, analyse de la séquence inverse complémentaire et réalisation de post-traitements pour résoudre le problème des ombres des CDS.

Si *EcoParse* n'a pas eu le succès de *GeneMark* c'est, entres autres, parce qu'un seul brin de la séquence est modélisé. Par ailleurs, la modélisation des gènes dans *EcoParse* ne prend pas en compte l'hétérogénéité de composition en oligonucléotides des CDS, plus particulièrement l'existence possible d'un groupe de gènes de composition atypique (riches en A+T), qui sont généralement mal prédits [Borodovsky *et al.*, 1995].

En 2003, T. Larsen et A. Krogh publient *EasyGene*, un programme de prédiction de gènes procaryotes fondé sur une architecture HMM ([Larsen & Krogh, 2003] Tab. 3.3 p. 91). Cette architecture représente un modèle *nul* et un modèle de gène sur un seul brin de la séquence. Le modèle nul est constitué du sous-modèle de *fond* (*background*) qui décrit la composition générale du génome pour capturer l'intergénique, et du sous modèle d'*ombres* (*shadows*) qui décrit un codon

complémentaire pour reconnaître les régions codantes du brin complémentaire. Le modèle de *gènes* est composé de sous-modèles pour le RBS, le codon d'initiation, les trois bases qui suivent le codon d'initiation, les trois types de codant (trois codons sens pour chaque type de codant), les trois bases qui précèdent le codon de terminaison, le codon de terminaison et enfin les six bases qui suivent le codon de terminaison. Le contexte des codons d'initiation et de terminaison est donc ici modélisé.

La phase d'apprentissage est automatique mais complexe, et pas complètement *ab initio*. En effet, l'extraction automatique de jeux de séquences de grande qualité repose sur la recherche de similitudes entre le produit des CDS de longueur supérieure à 120 pb et les séquences de la banque Swiss-Prot (méthode extrinsèque). Dans une seconde étape, les paramètres du modèle HMM sont estimés à l'aide de l'algorithme *Forward-backward*.

La phase de reconnaissance utilise aussi l'algorithme *Forward-backward* qui permet de calculer la probabilité qu'un gène démarre en une position donnée, sous le modèle utilisé. Cette approche est préférable à celle de l'algorithmne de Viterbi, car chaque gène prédit est accompagné d'une valeur de confiance. L'architecture de ce HMM ne modélise qu'un seul gène sur la séquence (le modèle ne boucle pas sur lui-même) ; la probabilité calculée n'a donc pas de réelle signification biologique : elle correspond à la probabilité que cette unique CDS soit en une position donnée. C'est pourquoi A. Krogh construisit un test statistique visant à rejeter l'hypothèse *il n'y a pas du tout de gène dans la séquence* par rapport à l'hypothèse *il y a un gène qui commence à une position donnée*. Ce test permet de calculer une E-value (appelée R) qui correspond au nombre de CDS artefactuelles *de n'importe quelle longueur* prédites avec un score supérieur à celui de la CDS potentielle, dans une séquence aléatoire de longueur 1 Mbp (une séquence ne contenant *a priori* pas de « vrais » gènes). La E-value est calculée de façon à ce que le taux de faux-positifs soit indépendant de la longueur des gènes prédits, ce qui revient à favoriser les longues CDS au détriment des petites. En effet, étant donné la distribution des longueurs des ORF artefactuelles dans une séquence, imposer un taux de faux-positifs indépendant de la longueur des gènes prédits revient à creuser l'écart entre la E-value des petites CDS et celle des grandes CDS[22]. Une petite CDS contenant moins d'information qu'une grande CDS, elle aura tendance à avoir une plus mauvaise E-value. La phase de reconnaissance effectuée sur le brin direct, puis sur le brin inverse, calcule la E-value de chaque gène (nombre attendu de faux-positifs de n'importe quelle longueur, prédits avec un score supérieur à celui du gène, dans 1 Mpb de séquence aléatoire). Toutes les CDS dont la E-value est inférieure à un certain seuil sont sélectionnées (la valeur seuil par défaut est égale à 2). La modélisation explicite du RBS, du contexte du codon d'initiation et l'apprentissage de ces modèles sur des séquences en $5'$ d'un jeu de CDS dont les codons d'initiation ont été déterminés avec certitude (méthode extrinsèque), permet d'ajuster le codon d'initiation des CDS prédites et ainsi de minimiser les chevauchements liés au choix du codon le plus en $5'$. Si un gène présente plusieurs codons d'initiation possibles et

[22]Dans le cas d'une petite CDS, une E-value qui mesure le nombre de « fausses » CDS d'une certaine longueur qui ont un score supérieur à certain seuil, va être nettement inférieure à la E-value qui mesure le nombre de « fausses » CDS de n'importe quelle longueur qui ont un score supérieur à ce même seuil ; alors que dans le cas d'une grande CDS, la E-value dépendante de la longueur ne sera que légèrement inférieure à la E-value indépendante de la longueur.

que les CDS correspondantes ont une E-value inférieure au seuil, *EasyGene* permet de les ordonner selon leur significativité et ainsi d'aider au choix du « vrai » codon d'initiation.

EasyGene est l'un des meilleurs programmes pour la prédiction des codons d'initiation. Il est possible de détecter des gènes ne possédant pas de RBS standard (gènes compactés en opéron ou qui présentent un décalage du cadre de lecture) *via* la transition directe dans le graphe, du modèle nul au codon d'initiation sans passer par le RBS. A. Krogh ignore et résout à la fois le problème des chevauchements, puisque chaque gène est évalué indépendamment des chevauchements qu'il peut avoir avec d'autres gènes (les chevauchements ne sont pas explicitement modélisés). Le fait de ne modéliser qu'un seul brin n'est pas gênant pour la prédiction, puisque les régions d'ombres sont modélisées. Trois sous modèles de codant permettent de décrire l'hétérogénéité des CDS, mais que se passerait-il si elle devait être décrite par quatre types de composition ?

Programmes *RHOM* et *SHOW* Le programme *RHOM*, développé par F. Muri, permet de segmenter une séquence en régions homogènes selon la composition locale en oligonucléotides (Tab. 3.3 p. 91). Le modèle HMM est très simple, les seuls paramètres attendus par le programme étant le nombre d'états cachés q, et l'ordre des chaînes de Markov des différents états m. La phase d'apprentissage des paramètres, et la phase de reconnaissance des régions homogènes, utilisent l'algorithme Forward-backward. Pour un modèle donné, un jeu de paramètres et un nombre d'itérations donnés, *RHOM* renvoie à la fois l'estimation des paramètres et les probabilités des régimes cachés en chacune des positions de la séquence. Pour chaque état caché considéré, les probabilités du régime, en fonction de la position sur la séquence, sont représentées graphiquement afin de visualiser les plages homogènes détectées par *RHOM*. A titre d'exemple, l'utilisation d'un modèle $M1-M3$ à trois états cachés, sur le chromosome de *B. subtilis*, a permis de mettre en évidence trois groupes de gènes qui ont une signification biologique : les gènes transcrits dans le sens direct, les gènes transcrits dans le sens inverse, et les gènes de composition atypique, généralement riches en nucléotides A+T. Ce dernier groupe contient aussi les régions intergéniques qui ont tendance à être riches en bases A+T [Nicolas *et al.*, 2002]. Le programme *RHOM* peut être considéré comme une méthode alternative à la méthode AFC – K-means (voir p. 121). En effet, il est possible d'attribuer un numéro d'état caché à chacun des gènes annoté en fonction de la segmentation du chromosome définie par *RHOM* et ainsi de former des groupes de gènes. Par exemple, on peut définir que les classes I, II et III, contiennent respectivement, les gènes transcrit dans le sens direct, les gènes transcrit dans le sens inverse et les gènes de composition atypique.

Le logiciel *SHOW*, successeur de *RHOM*, a le même objectif qu'*EasyGene* (Tab. 3.3 p. 91) dans le sens où ils abordent tous deux le problème de l'évaluation statistique d'une prédiction de gène dans un cadre HMM, évaluation résolue avec l'algorithme Forward-backward. *SHOW* se distingue cependant d'*EasyGene* par le fait qu'il permet à l'utilisateur de concevoir des HMM pour la détection de gènes à partir d'un fichier de configuration, et d'estimer leurs paramètres *ab initio* sans avoir besoin d'extraire des jeux de CDS de confiance en utilisant des similitudes au niveau protéique. La description de l'architecture HMM est modulaire, le format du fichier de configuration

autorisant une définition souple du modèle et des contraintes de l'estimation (paramètres fixés, couplés, mots interdits, etc). Il devient alors possible d'utiliser *SHOW* non seulement pour faire de la prédiction de gènes, mais aussi pour segmenter un génome ou des CDS en régions homogènes. Le programme *SHOW* propose plusieurs modèles déjà implémentés, par exemple celui permettant la détection de gènes procaryotes. A l'inverse du modèle proposé dans *EasyGene*, celui de *SHOW* décrit tous les gènes sur les deux brins du chromosome (HMM complet bouclé, comme dans le programme *GeneMark.hmm* ; voir p. 114). Ce modèle est complexe car il utilise des HMM avec un grand nombre d'états cachés et de nombreuses contraintes sur les paramètres. A travers ces contraintes, les états cachés reflètent les différents régimes observés le long d'un génome : régions intergéniques longues ou courtes, ARN fonctionnels, motifs RBS, codons d'initiation, plusieurs groupes de gènes codants, codons de terminaison, chevauchements longs et courts entre deux gènes de même sens, etc. A l'exception des régions intergéniques, les sous-modèles existent symétriquement pour le brin direct et pour le brin inverse. Enfin, à l'inverse d'*EasyGene*, la probabilité calculée par l'algorithme Forward-backward sur ce modèle a une signification immédiatement interprétable : elle mesure directement la probabilité qu'il y ait un gène en une position donnée de la séquence. Il suffit alors de prédire tous les gènes dont la probabilité dépasse un certain seuil, et ce, quelque soit leur longueur. Lorsque le choix du codon d'initiation d'une CDS est ambigu, *SHOW* permet d'en prédire plusieurs et d'attribuer, à chacun d'eux, une probabilité de démarrer effectivement le gène. Ainsi, *SHOW* permet de gérer les problèmes fondamentaux de recouvrement entre CDS, de détection du « vrai » codon d'initiation, de détection des petites CDS et des CDS de composition atypique (riches en A+T).

Les programmes *EasyGene* et *SHOW* permettent tous deux de donner une valeur de confiance dans les prédictions de gènes grâce à l'algorithme Forward-backward. Celle de *SHOW* est directement la probabilité calculée par Forward-backward alors que celle d'*EasyGene* est une E-value qui nécessite un calcul compliqué. Le choix simple (et surprenant) d'un modèle non bouclé, à un gène et sur un seul brin de la séquence, dans *EasyGene* est très différent du choix complexe (et réaliste) d'un modèle bouclé de gènes, sur les deux brins de la séquence, dans *SHOW*. La phase d'apprentissage est automatique et intrinsèque dans le cas *SHOW*, mais extrinsèque dans le cas de *EasyGene*. Enfin, le programme *SHOW* est distribué librement (un serveur web permet d'accéder à *EasyGene*). On comprend mieux les choix de T. Larsen quand on s'aperçoit, qu'en pratique, les résultats d'*EasyGene* sont de qualités équivalentes à ceux de *SHOW* [Nicolas, 2003].

Programme *Frame-by-frame* La méthode de détection de gènes *Frame-by-frame* développée par A. Shmatkov et M. Borodovsky ([Shmatkov *et al.*, 1999] Tab. 3.3 p. 91) a pour objectif d'améliorer la précision des prédictions de gènes données par *GeneMark.hmm* (voir p. 114). Cela est possible grâce aux deux caractéristiques principales du HMM de *Frame-by-frame*. D'une part, chacune des six phases de lecture est analysée indépendamment, puis le décodage final est produit en regroupant les résultats de ces six analyses, autorisant la prédiction de CDS chevauchantes. D'autre

part, la modélisation détaillée des régions qui encadrent le codon d'initiation[23] permet d'ajuster au mieux la position du codon d'initiation des gènes prédits.

L'architecture HMM décrit les régions codantes d'une phase de lecture particulière, examinée dans sa globalité. Il n'est donc plus nécessaire de modéliser ni les différentes phases de la séquence, ni les chevauchements (qui par définition se produisent toujours entre deux CDS de phase différente). Un état caché décrit un *triplet*, et non un nucléotide ou une chaîne de nucléotides de taille variable. L'alphabet considéré est donc constitué de soixante-quatre *triplets* (et non pas de 4 lettres). Les transitions entre états cachés suivent la logique de l'organisation des gènes procaryotes :

1. le *codon* d'initiation est émis par un état caché S
2. les n *codons* qui suivent sont modélisés par n états C_i, puis les *codons* restant de la CDS sont émis par un état bouclé C
3. le *codon* de terminaison est émis par un autre état E
4. il est suivi par l'état T bouclé qui modélise les *triplets* non-codants, enfin les m *triplets* qui précédent le *codon* d'initiation suivant sont émis par m états $T_m \dots T_i \dots T_1$

Les états C_i et T_i permettent de décrire en détail les régions de part et d'autre du codon d'initiation. Les états T et T_i modélisent les régions qualifiées de *trous*, c'est-à-dire des régions comprises entre deux gènes codés dans la même phase. Le modèle complet a une forme circulaire et certaines transitions qui coupent le cercle autorisent le passage direct de l'état codon de terminaison à l'état codon d'initiation par exemple (sans passer par les états de *trous*).

La phase d'apprentissage est réalisée sur des séquences génomiques annotées afin d'estimer les probabilités de transition entre états cachés, et les probabilités d'émission des *triplets* à partir de leurs fréquences (hypothèse nulle d'apparition indépendante et aléatoire des *triplets*). Par exemple les probabilités d'émission des codons de l'état C sont proportionnelles aux fréquences relatives des codons de la table de l'usage du code génétique établies sur les CDS annotées ($F_R(abc)$).

La phase de reconnaissance lit chacune des six phases d'une séquence nucléique, *triplet* par *triplet*. L'algorithme de Viterbi permet de trouver le chemin le plus probable des régions codantes sur chacune des six phases de lecture. La superposition des six résultats autorise les chevauchements et les inclusions. Un post-traitement élimine ensuite les CDS incluses puisqu'il n'y a pas d'évidence expérimentale pour de tels cas de figure.

Le programme *Frame-by-frame* a une sensibilité moins élevée que celle de *GeneMark.hmm* ce qui est probablement dû à l'absence de modélisation des gènes de composition atypique. En revanche, il a une meilleure spécificité et une meilleure précision dans le choix du *codon* d'initiation de la traduction.

[23] m triplets en $5'$ du codon d'initiation et n codons en $3'$.

Méthodes fondées sur les modèles semi-markoviens cachés

Dans les HMM, le caractère markovien des transitions entre les états implique que la longueur des plages dans un état (encore appelé *durée* de séjour) suit une loi géométrique[24] (les petites plages sont plus probables que les grandes plages). Or, ceci n'est pas toujours vérifié dans les séquences biologiques.

Si l'on construit un histogramme des longueurs des séquences non-codantes procaryotes, la distribution suit bien une loi géométrique (de longueur moyenne 150 pb, ce qui n'est absolument pas le cas des séquences génomiques Eucaryotes).

En revanche, la distribution des longueurs des « vraies » CDS suit plutôt une loi Γ discrétisée qu'une loi géométrique, indiquant que les très petites et les très grandes CDS sont toutes deux peu probables (la taille moyenne d'une CDS dans un génome bactérien est de 900 pb [Lukashin & Borodovsky, 1998]).

Afin de modéliser les CDS de manière plus réaliste, il faudrait donc quitter la plage codante en moyenne au bout de 900 pb. Dans ce cadre, la solution des *modèles semi-markoviens cachés* (*Hidden Semi-Markov Models* (*HSMM*) encore appelés *Generalised HMM*, *Explicit Duration State HMM*), extension des HMM, semble plus adaptée, car elle permet de modéliser explicitement la durée des séjours dans les états cachés en utilisant, par exemple, une loi Γ discrétisée pour modéliser la taille des CDS, et une loi exponentielle pour modéliser la taille des séquences non-codantes. La loi de probabilité d'apparition d'un nucléotide décrivant une région homogène peut être définie par n'importe quel type de modèle statistique : un modèle de Bernoulli $M0$, une table d'usage des codons, un modèle de chaînes de Markov inhomogène Mm_3, un IMM, un réseau de neurones, etc. Ainsi les HSMM permettent aussi de modéliser en un état ce qui nécessite plusieurs états dans le cadre des HMM. La principale contrepartie de l'utilisation des HSMM est d'augmenter considérablement le coût des calculs. Evidemment, un HSMM peut être équivalent à un HMM standard si on définit la loi de durée de séjour dans un état par la loi géométrique du HMM standard [Rabiner, 1989].

Programme *GeneMark.hmm* 1.0 Dans le cas du programme *GeneMark.hmm* (Tab. 3.3 p. 91, [Lukashin & Borodovsky, 1998]), un état caché ne représente pas un nucléotide mais une chaîne de nucléotides d'une certaine longueur (un modèle = un état). Le modèle HSMM utilisé compte neuf états cachés : non-codant de longueur n, codon d'initiation direct et inverse, codant direct typique de longueur i et codant inverse typique de longueur k, codant direct atypique de longueur j et codant inverse atypique de longueur m, codon de terminaison direct et inverse. Le modèle complet est bouclé, il représente donc les gènes sur les deux brins de la séquence. Ce HSMM repose sur les modèles de chaînes de Markov de *GeneMark*, c'est-à-dire des chaînes de Markov inhomogènes (Mm_3), pour les modèles codants typique et atypique sur chacun des deux brins, et des chaînes de Markov homogènes (Mm) pour le non-codant. Ainsi, on peut modéliser une région

[24] Une loi géométrique est une loi exponentielle discrétisée.

codante de différentes manière : trois états cachés bouclés, représentant chacun un nucléotide sous le modèle Mm (*EasyGene* ou *SHOW*), $61*3$ états cachés bouclés représentant les trois nucléotides des soixante et un codons sous le modèle $M0$ (*EcoParse*), un seul état caché bouclé représentant un codon sous le modèle $M0$ (*Frame-by-frame*), un seul état caché non bouclé représentant une séquence de nucléotides, dont la longueur est modélisée explicitement par la durée de séjour dans l'état codant, sous le modèle Mm_3.

Dans *GeneMark.hmm*, il n'y a pas de phase d'apprentissage par l'algorithme Forward-backward. La lourdeur de la phase d'apprentissage automatique d'un HMM et de surcroît d'un HSMM, est ainsi évitée. Les paramètres des modèles codant et non-codant de ce HSMM regroupent les probabilités d'émission des nucléotides au sein d'un état caché et les probabilités de séjour dans un état caché, ce qui permet de générer une séquence d'une certaine longueur pour chacun de ces modèles. Pour les probabilités d'émission des nucléotides au sein des modèles codant et non-codant, on peut utiliser les matrices de transitions générées par le programme *GeneMark-Genesis*. La probabilité $p_u(d)$ qu'un état caché u ait une durée d est définie par approximation analytique de la distribution des probabilités des longueurs des régions codantes (resp. non-codantes) extraites à partir des annotations du chromosome d'*E. coli* K-12 (distribution gamma pour le codant et distribution exponentielle pour le non-codant [Lukashin & Borodovsky, 1998]). D'autres paramètres sont estimés à partir des statistiques des annotations du génome d'*E. coli* K-12 comme les probabilités d'émission des codons d'initiation sous le modèle $M0$ estimées à partir de la fréquence des codons d'initiation ($F_R(ATG) = 0,905$; $F_R(GTG) = 0,090$; $F_R(TTG) = 0,005$). Les probabilités de transitions entre états cachés sont peu nombreuses car le nombre de transitions autorisées dans le graphe est faible (*non codant* $\rightarrow$ *initiation direct* $\rightarrow$ *codant direct* $\rightarrow$ *terminaison direct* $\rightarrow$ *non codant*). Les valeurs des probabilités de transition de l'état non-codant à l'état codant typique et atypique sont estimées à partir des fréquences des gènes typiques (resp. atypiques) dans le génome d'*E. coli* K-12 (0,85 (resp. 0,15) ; [Médigue *et al.*, 1991, Lawrence, 1997]).

La phase de reconnaissance utilise une extension de l'algorithme récursif de Viterbi adaptée au HSMM, qui maximise la probabilité conditionnelle ($\mathbb{P}_\theta(S_1^n = s_1^n \mid X_1^n = x_1^n)$) de trouver, en segmentant la séquence d'ADN observée x_1^n, la meilleure séquence s_1^n* d'états fonctionnels. *GeneMark.hmm*, qui interdit la possibilité de recouvrements entre gènes adjacents, a tendance à prédire les gènes impliqués dans des chevauchements plus courts qu'ils ne le sont en réalité (forcément du côté $5'$). Il ne modélise pas non plus les RBS et ne résout donc pas le problème du choix du codon d'initiation lors de la phase de reconnaissance. Aussi, pour palier à ce dernier problème, une procédure de post-traitement de recherche du motif RBS permet d'affiner les prédictions de Viterbi. Le codon d'initiation est repositionné si le score d'un des RBS candidats associés à un codon d'initiation alternatif excède un certain seuil. Le modèle probabiliste du RBS est une matrice de fréquences de nucléotides spécifiques de la position, dont les paramètres sont dérivés à partir d'alignements multiples par échantillonnage de Gibbs, et de séquences localisées en $5'$ des codons d'initiation annotés.

GeneMark et *GeneMark.hmm* ont finalement des propriétés complémentaires. Les gènes impli-

qués dans des chevauchements sont manqués par *GeneMark.hmm*, mais peuvent être correctement prédits par *GeneMark*. Inversement, les gènes de composition atypique sont manqués par *GeneMark*, mais peuvent être prédits correctement par *GeneMark.hmm*. Il est donc intéressant, dans le cadre d'une annotation fine (où l'on cherche à minimiser le nombre de faux-négatifs), de combiner les prédictions de *GeneMark.hmm* et de *GeneMark*.

Programmes *GeneMark.hmm* 2.0 et *GeneMarkS* Ces deux derniers programmes sont présentés ensemble car *GeneMarkS* utilise *GeneMark.hmm* 2.0 (qui repose sur les modèles de *GeneMark* ; [Besemer *et al.*, 2001] TAB. 3.3 p. 91). De plus [Besemer & Borodovsky, 1999], un autre programme sans nom, permet, par une approche heuristique, de construire des matrices d'émission de nucléotides qui peuvent être utilisées à la fois par *GeneMark* et par *GeneMark.hmm* (et donc par *GeneMarkS*). J. Besemer et M. Borodovsky ont remarqué que la prédiction des gènes des différents groupes homogènes dans leur usage des codons synonymes ne nécessitait pas forcément l'utilisation d'une matrice de probabilités de transition pour chaque groupe de gènes.

Par exemple, dans le cas des trois classes d'*E. coli* K-12 [Borodovsky *et al.*, 1995], les gènes de classe II (hautement exprimés) peuvent être précisément prédits juste avec la matrice des gènes de classe I (qui regroupe la majorité des gènes). La matrice de classe III qui regroupe les gènes atypiques (A+T riches) est celle qui prédit le plus de gènes des trois classes. Autrement dit, cette matrice a en moyenne une meilleure sensibilité pour prédire un gène de n'importe quelle classe que les deux autres matrices (la matrice I prédit mal les gènes de classe III) ; mais bien entendu, chaque matrice est la plus spécifique de sa classe (la meilleure matrice pour prédire les gènes de classe I reste la matrice I). Partant des seuls paramètres de la composition en nucléotides et du pourcentage en G+C de la séquence à analyser, J. Besemer a mis au point une heuristique pour construire un modèle de chaîne de Markov 3-périodique des régions codantes, et un modèle de chaîne de Markov homogène des régions non-codantes. Il suffit d'un fragment de 400 pb pour estimer ces paramètres.

La méthode consiste à utiliser des fonctions linéaires qui permettent de relier d'une part les fréquences globales en nucléotides aux fréquences en nucléotides aux trois positions des codons, et d'autre part le pourcentage en G+C aux fréquences en acides aminés. Ces fonctions linéaires ont été définies à partir d'une étude portant sur les annotations de dix-sept génomes bactériens complets. Pour établir la relation entre la fréquence globale d'un nucléotide et sa fréquence au trois positions des codons, J. Besemer a représenté pour chaque nucléotide, en abscisse sa fréquence globale et en ordonnée ses trois fréquences positionnelles dans les dix-sept jeux d'annotation [Besemer & Borodovsky, 1999]. A partir de ces quatre graphiques, il a déduit par régression linéaire les $3*4$ corrélations existant entre taux global et taux positionnel. De même, pour établir la relation entre le pourcentage en G+C et la fréquence en acide aminé, il a représenté pour chaque acide aminé, en abscisse le pourcentage en G+C, et en ordonnée la fréquence de l'acide aminé dans les dix-sept jeux d'annotation. Parmi les vingt acides aminés, seuls dix ont une fréquence qui varie significativement avec le pourcentage en G+C, l'étude n'a donc portée que sur ces dix acides aminés (alanine, glycine, proline, arginine, phénylalanine, isoleucine, lysine, asparagine, tyrosine et valine). A partir

de ces dix graphiques, il a déduit par régression linéaire les dix équations définissant les corrélations entre fréquence des acides aminés et pourcentage global en G+C. Ces fonctions linéaires, définies une fois pour toute, permettent alors de construire une table d'usage du code génétique spécifique de la composition en nucléotides et du pourcentage en G+C de la séquence donnée en entrée. Les valeurs initiales de fréquence d'occurrence de chacun des soixante et un codons $F_I(abc)$ sont obtenues par le produit de trois fréquences positionnelles des nucléotides correspondant. La fréquence relative d'un codon est calculée par le produit de la fréquence relative de l'acide aminé correspondant et de la fréquence relative du codon synonyme. Par exemple pour le codon GCT de l'alanine, on a : $F_R(GCT) = F_R(alanine) \times (F_I(GCT)/(F_I(GCA)+F_I(GCC)+F_I(GCG)+F_I(GCT)))$. Pour construire la matrice de transition, on utilise cette table pour les modèles tri-périodiques codant direct et inverse, et les fréquences globales des nucléotides pour le modèle non-codant d'ordre 0. Cette table permet de calculer les paramètres du modèle de chaînes de Markov 3-périodique d'un pseudo ordre 2 des régions codantes : la probabilité d'apparition d'un nucléotide en troisième position des codons dépend des deux nucléotides qui précèdent ($M2$), celle en deuxième position dépend du nucléotide qui précède ($M1$) et celle en première position dépend uniquement de la fréquence en nucléotides à la première position des codons ($M0$). Autrement dit, un modèle de pseudo ordre 2 équivaut à un modèle à dépendance variable en fonction de la position dans le codon.

Finalement, J. Besemer compare la matrice heuristique d'*E. coli* K-12 avec les matrices des trois classes de gènes d'*E. coli* K-12 au moyen de l'indice de distance de Kullback-Liebler calculée entre ces différentes matrices (voir p. 98). Il apparaît que c'est la matrice de classe III qui est la plus proche de la matrice heuristique. La matrice heuristique permet donc, comme la matrice de classe III, de prédire un maximum de gènes des trois classes et en particulier, le groupe des gènes atypiques qui sont les plus difficiles à prédire de par leur composition A+T riche qui les rapproche des séquences non-codantes. La matrice heuristique possède cependant trois inconvénients :

1. elle n'a pas été construite à partir de séquences d'ADN réelles (les valeurs sont approximatives)
2. l'ordre de Markov utilisé est au maximum de 2
3. ce modèle fait l'hypothèse que deux génomes ayant le même pourcentage en G+C auront le même usage des codons

Comme nous allons le voir ci-dessous, ce type de matrice peut être tout à fait adapté pour initialiser un processus de construction de matrices (utilisée alors comme prématrice) et/ou pour servir de matrice atypique.

La version de *GeneMark.hmm* 2.0 comporte essentiellement deux améliorations (Tab. 3.3 p. 91). La nouvelle architecture du HSMM intègre un modèle décrivant la région en $5'$ du codon d'initiation, afin d'améliorer la précision de la prédiction du « vrai » codon d'initiation des gènes. Le modèle RBS est composé de deux sous modèles (deux états) :

1. le motif SD de 6 pb (matrice de fréquences des nucléotides spécifique de la position)
2. la région qui sépare le site SD du codon d'initiation (matrice de fréquences des nucléotides dans les séquences non-codantes et probabilité de séjour calculée grâce à la distribution des

probabilités des longueurs des espaceurs)

Un gène est donc défini par l'ensemble des états : SD, espaceur, codon d'initiation, région typique ou atypique, codon de terminaison. Une seconde amélioration concerne la modélisation des chevauchements de tout type entre deux gènes (chevauchement court ou long, entre une CDS et le RBS d'une CDS juxtaposée, entre deux CDS même sens, sens contraire, etc.).

GeneMarkS, fondé sur *GeneMark.hmm* 2.0, apprend et reconnaît les gènes, en particulier le « vrai » codon d'initiation, par une procédure itérative automatique et *ab initio*. Les principales étapes de la procédure *GeneMarkS* peuvent être résumées de la façon suivante :

1. Lors de l'initialisation,
   (a) le pourcentage en G+C et la fréquence globale en nucléotides, calculés à partir d'une séquence d'ADN anonyme, servent à dériver les modèles heuristiques codant et non-codant
   (b) la version *allégée* de *GeneMark.hmm* 2.0, qui ne modélise ni les gènes atypiques, ni les RBS, utilise alors cette prématrice afin d'identifier les régions codantes et non-codantes dans chacune des trois phases, sur chacun des deux brins de la séquence
   (c) on extrait un jeu de séquences non-codantes de 25 pb en $5'$ des codons d'initiation des régions codantes précédemment prédites, sur lequel un alignement multiple par échantillonnage de Gibbs est réalisé. Cet alignement sert à définir les paramètres du sous-modèle SD et du sous modèle espaceur.
   (d) Les itérations alternent deux étapes :
      i. construction de la matrice *pseudo native* puisqu'elle est dérivée à partir des CDS et séquences non-codantes classées *in silico* par *GeneMark.hmm* 2.0 (et non à partir de séquences vérifiées expérimentalement ou par des annotateurs) et construction de la matrice des sous modèles RBS dérivés à partir de l'alignement multiple par échantillonnage de Gibbs de séquence en $5'$ des CDS prédites par *GeneMark.hmm* 2.0
      ii. prédiction de gènes avec la version *complète* de *GeneMark.hmm* 2.0 en utilisant la prématrice heuristique (déterminée lors de l'initialisation) pour les modèles codants atypiques, la matrice pseudo native pour les modèles codants typiques et la matrice RBS (calculées à l'étape précédente).
   (e) La convergence est atteinte soit quand le jeu de gènes est à 99% identique à celui de l'itération précédente, soit quand le pourcentage d'identité fluctue autour d'une valeur suffisamment haute.

Les résultats de *SHOW*, d'*EasyGene* et de *GeneMarkS* sont de qualité équivalente [Nicolas, 2003]. Le programme *GeneMarkS* prédit mieux les codons d'initiation que *GLIMMER* ; il est moins sensible mais plus spécifique que *GLIMMER* [Besemer *et al.*, 2001].

Conclusions des méthodes de prédiction de gènes Dans leur publication sur *EasyGene* [Larsen & Krogh, 2003], les auteurs comparent les résultats de sept programmes de prédiction de gènes : *EasyGene*, *GLIMMER*, *Orpheus*, *GeneMark*, *GeneMark.hmm*, *GeneMarkS* et *Frame-by-frame*. Ils en concluent que les programmes de prédiction de gènes les moins spécifiques (qui génèrent le plus de faux-positifs) sont *Orpheus* et *GLIMMER*, puis viennent *GeneMark.hmm* et *GeneMarkS*, et enfin *GeneMark*, *Frame-by-frame*, et *EasyGene*. Les programmes les moins sensibles (qui génèrent le plus de faux-négatifs) sont *GeneMark*, *Frame-by-frame*, *GeneMarkS*, puis *Orpheus*, et enfin *GeneMark.hmm*, *EasyGene* et *GLIMMER*. Si on prend en compte l'ensemble des critères : sensibilité et spécificité des programmes, prédiction exacte du codon d'initiation, prédiction correcte des petits gènes et des gènes impliqués dans des chevauchements, *EasyGene* arrive en tête.

Nous avons vu que, seuls les modèles HMM utilisés dans les programmes *GeneMarkS*, *SHOW* et *EasyGene*, sont capables de modéliser explicitement et efficacement les motifs RBS pour améliorer la prédiction des gènes (trouver le codon d'initiation correct qui généralement minimise les problèmes de chevauchement). Il est par ailleurs intéressant de modéliser les différents types de composition des CDS, ce qui est réalisé avec les méthodes fondées sur les HMM (*GeneMarkS*, *SHOW* et *EasyGene*). Cependant, ces méthodes modélisent difficilement les possibilités de recouvrement entre CDS (dans *EasyGene* ils sont ignorés, dans *SHOW* et *GeneMarkS*, les auteurs alourdissent l'architecture HMM en les modélisant). Les calculs mis en œuvre dans les méthodes de prédiction de gènes par HMM sont beaucoup plus lourds que ceux des méthodes par chaînes de Markov, puisque la phase d'apprentissage et la phase de reconnaissance reposent sur des algorithmes itératifs de maximisation de la vraisemblance, qui permettent à la fois d'estimer les paramètres et de segmenter la séquence dans le cas des modèles à données incomplètes. Finalement, au cours du processus de prédiction de gènes codants, il apparaît que les phases indispensables à une prédiction correcte sont, par ordre d'importance : le choix du modèle probabiliste, la phase d'apprentissage permettant d'estimer les paramètres du modèle (la construction des jeux de séquences codantes et non-codantes pour l'apprentissage), la phase de reconnaissance qui doit donner une signification statistique aux résultats (calcul d'un potentiel de codage), et enfin la phase optionnelle de post-traitement qui permet d'affiner les résultats.

3.4 Prédiction de décalages du cadre de lecture

Généralement, une CDS correspond à une séquence nucléique orientée de $5' \rightarrow 3'$, débutant par un codon d'initiation et finissant au niveau du premier codon de terminaison en phase, rencontré. Cependant, il peut arriver que la CDS soit fragmentée. Le fragment le plus en $5'$ commence par le « vrai » codon d'initiation, celui le plus en $3'$ se finit par le « vrai » codon de terminaison (pas nécessairement en phase avec le codon d'initiation) et les longueurs des différents fragments sont des multiples de trois. Ces décalages du cadre de lecture peuvent correspondre soit à une erreur de séquence au cours du séquençage, soit à un saut de phase authentique. Ce dernier cas peut empêcher l'expression du gène (*pseudogène*) ou réguler son expression (décalage du cadre de lecture

programmé pour réguler la traduction du gène). Les décalages du cadre de lecture sont provoqués par l'insertion ou la délétion d'un nombre de nucléotides non multiple de trois (artefact de séquençage ou mutation). On peut étendre la définition de *frameshifts* à d'autres types d'anomalie, par exemple : une substitution générant un codon stop en phase (deux fragments de CDS sans décalage de phase), une délétion d'un fragment de CDS (CDS partielle), une IS insérée dans une CDS (l'IS peut être dans le même sens que la CDS ou dans le sens opposé), etc.

La recherche de *frameshifts* dans les séquences génomiques consiste à définir les positions du chromosome où l'on observe des décalages du cadre de lecture. Il existe deux types d'approche : intrinsèque et extrinsèque. Dans le cadre des approches intrinsèques, le programme *FSED*, repose sur l'usage du code génétique des gènes de l'organisme étudié et identifie les ruptures brutales dans le profil d'usage des codons au sein des CDS annotées [Fichant & Quentin, 1995]. Cette méthode est particulièrement adaptée à la détection de *frameshifts* dits *compensés*. Ce cas de figure est défini par un premier décalage du cadre de lecture compensé plus loin par un second, qui rétablit la région codante dans sa phase initiale. En pratique, on observe que le fragment en $5'$ inclut le fragment en $3'$. Le fragment en $5'$ possède le « vrai » codon d'initiation et le « vrai » codon de terminaison, mais la courbe de probabilité de codage chute au niveau de la région d'inclusion (partie artificielle du fragment en $5'$). Le fragment inclus est donc mis en évidence par son potentiel de codage. Deux décalages du cadre de lecture successifs sont donc à l'origine du *frameshift compensé*. Ce double événement, plutôt rare, ne révèle pas toujours une erreur de séquence [Rojas *et al.*, 2003]. Dans la même catégorie d'approche, la méthode *ProFED* [Médigue *et al.*, 1999b] permet de combiner les résultats d'une recherche par signal et d'une recherche par contenu sur la séquence génomique pour détecter deux types de *frameshifts* :

1. un codon d'initiation de CDS tardif par rapport au démarrage de la courbe de codage (dont le cas extrême est un potentiel de codage significatif sans CDS)
2. des CDS chevauchantes avec un potentiel de codage significatif (dont le cas extrême est un *frameshift* compensé)

La recherche par signal (*prokov_orf*) permet de définir le jeu des CDS maximales d'une certaine longueur dans les six phases de la séquence. La recherche par contenu (*prokov_curve*) permet de mesurer le potentiel de codage le long des six phases en utilisant une matrice de transitions. Pour détecter les *frameshifts* de type *probabilité de codage sans CDS*, il faut d'abord créer, à partir du jeu de CDS, six vecteurs de CDS ayant, pour longueur, celle de la séquence génomique (0 quand on se situe entre deux CDS et 1 quand on est dans une CDS). De même, on définit six vecteurs de probabilités (à chaque position de la séquence correspond une valeur de probabilité de codage P_c). Puis, il suffit de parcourir le vecteur de probabilités et le vecteur de CDS pour chaque phase, et de repérer les régions d'une certaine longueur (au minimum 24 pb) qui ont à la fois un potentiel de codage significatif ($\mathbb{P}_\pi(COD_m \mid X_1^w = x_1^w) \geq 0,5$ pour les positions correspondantes dans le vecteur de probabilités) mais pas de CDS (0 dans le vecteur de CDS). Pour détecter les *frameshifts* de type *deux CDS chevauchantes dont le potentiel de codage est significatif*, il faut d'abord calculer la probabilité moyenne de codage associée à chaque CDS (Pc calculée par exemple par *compute_Pc*,

voir p. 207) ce qui permet de définir un jeu de CDS dont la probabilité moyenne de codage est supérieure à un certain seuil ($Pc \geq 0,25$). Puis, il suffit de repérer les chevauchements d'une certaine longueur (au minimum 36 pb) entre deux CDS de même sens. En fait, ce cas de figure correspond soit à un problème de réajustement du codon d'initiation, soit à un problème de décalage du cadre de lecture (les résultats de recherche de similitude permettront de trancher). Dans ce dernier cas, si les fragments sont en sens direct, alors le fragment le plus en $5'$ possède le « vrai » codon d'initiation mais pas le « vrai » codon de terminaison, et inversement, le fragment en $3'$ possède le « vrai » codon de terminaison mais pas le « vrai » codon d'initiation.

Aucune de ces deux méthodes ne permet de détecter le type d'anomalie où une CDS est interrompue par un codon de terminaison en phase, ce qui est en revanche possible avec une méthode de type extrinsèque (Fastx[25], Blastx). La méthode *Frame* [Brown *et al.*, 1998], est fondée sur une recherche de similitude (Blast2x) qui permet de comparer les six peptidiques traduites des six phases de lecture de la séquence génomique contre une banque de séquences protéiques. Les fragments protéiques sujets d'*une même entrée protéique* s'appariant significativement dans une même région génomique sont regroupés. Puis, pour chaque groupe (CDS candidate), on détecte la présence de décalage(s) du cadre de lecture et de codon(s) de terminaison en phase. Ce type d'approche est très efficace, mais d'une part, il est plus coûteux en ressources informatiques que le précédent, et d'autre part une similarité significative avec les données des banques doit exister pour qu'un décalage du cadre de lecture soit mis en évidence. L'idéal consiste donc à combiner les deux types d'approches, intrinsèque et extrinsèque, qui sont complémentaires.

3.5 Statistique descriptive de l'usage des codons dans les gènes

Les gènes codant les protéines, au cœur de ce travail, sont souvent analysés à plusieurs niveaux (nucléique et protéique). Les CDS correspondantes (définies par leur position de début et de fin, et par leur phase) peuvent être décrites par de nombreuses caractéristiques : leur taille, leur probabilité de codage, leur orientation par rapport à l'origine et au terminus de réplication (brin précoce ou tardif), leur implication dans un décalage du cadre de lecture, leur pourcentage en G+C moyen, leur usage des codons synonymes ou leur régulation transcriptionnelle. L'étude de l'utilisation des codons synonymes repose sur la mise en œuvre de méthodes d'analyse de données (statistique descriptive) : ces méthodes n'imposent aucune condition sur la distribution des données au départ, et ne modifient en rien les tendances présentes dans celles-ci. Ce sont des méthodes intrinsèques d'analyse par contenu. Nous allons voir ici, à la lumière de l'application relative à l'étude de l'usage du code génétique, les fondements des méthodes qui sont mises en œuvre dans le cadre de telles analyses.

[25] Fastx et Fasty autorisent les décalages du cadre de lecture dans les alignements (respectivement entre les codons et dans les codons) mais sont gourmands en ressources informatiques.

3.5.1 Indices

Deux approches sont possibles pour analyser des différences d'usage du code génétique d'un ensemble de gènes :

- Soit on utilise un indice qui synthétise en une seule valeur numérique l'usage relatif des 59 codons dans une CDS.
- Soit on décrit chaque CDS par un vecteur de 59 variables qui représentent l'usage de chaque codon dans une CDS (analyses multivariées, voir p. 121 et p. 133).

Le *GCskew* permet d'étudier des variations dans l'apparition des nucléotides le long des séquences nucléiques. Cet indice mesure la déviation entre les fréquences des nucléotides G et C : $GCskew = (G - C)/(G + C)$ [Rocha, 2000]. Il est généralement calculé à l'aide d'une fenêtre glissant le long du chromosome. Il permet souvent de mettre en évidence un biais mutationnel lié à la réplication et d'en déduire les positions de l'origine et du terminus de réplication.

D'autres indices permettent d'étudier les différences d'usage du code génétique entre deux gènes ou groupes de gènes. Certains indices mesurent la déviation entre l'usage du code génétique observé et un usage attendu. Trois hypothèses nulles H_0 sont utilisées :

1. l'usage du code génétique attendu est entièrement déterminé par le biais mutationnel
2. les codons sont utilisés de manière équiprobable
3. les codons *synonymes* sont utilisés de manière équiprobable (c'est l'hypothèse la plus usuelle car elle est simple et réaliste)

On peut citer, par exemple, les indices suivants : le GC_3[26], l'indice $P2$[27] [Gouy & Gautier, 1982], la fréquence relative des codons synonymes (Tab. 3.2 p. 89 ; cet indice est utilisé par le programme *Codon Preference*, [Gribskov *et al.*, 1984]), l'usage relatif des codons synonymes ou *Relative Synonymous Codon Usage* (*RSCU*) [Sharp *et al.*, 1986], et le calcul du nombre effectif de codons (*Effective Number of Codons* (EN_c) ; [Wright, 1990]).

Le RSCU est défini comme le rapport de la fréquence observée d'un codon dans un gène sur celle attendue, si l'usage de tous les codons synonymes codant le même acide aminé était uniforme (équiprobabilité des codons synonymes). Le RSCU se calcule alors de la manière suivante : $RSCU_{abc} = (Obs_{abc})/(Exp_{abc})$ où Obs_{abc} est le nombre observé d'occurrences du codon *abc*, et Exp_{abc} est le nombre attendu d'occurrences de codon *abc*. On a : $Exp_{abc} = (\sum N_{a'b'c'})/(N_{syn_{abc}})$ (Tab. 3.2 p. 89 et voir les tables d'usage des codons synonymes p. 230). Le RSCU, par rapport à la fréquence relative des codons synonymes (F_{RS}), tient compte du nombre de codons synonymes disponibles, ce qui facilite l'interprétation des résultats (la somme des RSCU est égale au nombre de codons synonymes 2, 3, 4 ou 6 ; le RSCU des codons uniques met et trp vaut systématiquement 1).

[26] L'indice du GC_3 mesure la fréquence relative des codons se terminant par G ou C dans un gène, à l'exception des codons uniques met, trp : $GC_3 = (N_{GC_3})/N$.

[27] L'indice $P2$ permet de calculer la proportion de codons qui sont conformes à la force de l'intermédiaire définie selon les règles d'interaction codon-anticodon. $P2 = (WWC + SSU)/(WWY + SSY)$ où $W = [A, U]$, $S = [G, C]$ et $Y = [C, U]$.

En conséquence, une valeur RSCU supérieure à 1 indique que le codon considéré est employé plus souvent qu'attendu, une valeur inférieure à 1 indique sa rareté et une valeur proche de 1 indique une absence de biais dans l'usage du codon.

D'autres indices mesurent les biais de codons à travers un jeu de codons préférés (optimaux), déterminé sur un ensemble de CDS d'apprentissage (ce sont souvent les gènes fortement exprimés, comme ceux codant les protéines ribosomiques). On peut citer, par exemple, l'indice du biais de codons (*Codon Bias Index*[28] (CBI), [Bennetzen & Hall, 1982]), la fréquence des codons optimaux (*Frequency of Optimal Codons* (F_{OP}) ; [Ikemura, 1985]) et l'indice d'adaptation des codons (*Codon Adaptation Index* (CAI) ; [Sharp & Li, 1987, Carbone *et al.*, 2003]). Le CAI est plus robuste d'un point de vue statistique que le CBI ou l'indice F_{OP}, car il n'est pas influencé par les effets de composition des génomes. L'indice CAI utilise un jeu de référence (gènes fortement exprimés d'un organisme donné) pour évaluer l'écart entre l'usage des codons synonymes des gènes du jeu de référence, et celui du gène considéré. Mathématiquement, le CAI d'un gène se calcule de la manière suivante :

$CAI = CAI_{obs}/CAI_{max}$

avec : $CAI_{obs} = \left(\prod_{k=1}^{N} RSCU_k\right)^{1/N}$, $CAI_{max} = \left(\prod_{k=1}^{N} RSCU_{k\ max}\right)^{1/N}$

et N le nombre de codons d'une CDS.

Les CAI_{obs} et CAI_{max} sont les moyennes géométriques des valeurs RSCU correspondant à tous les codons du gène (elles sont calculées à partir de l'ensemble de référence). CAI_{max} est la valeur maximale que peut prendre un gène de même composition en acides aminés. Pour chaque acide aminé on choisit, dans la table de référence, le codon (synonyme du codon observé) qui a le RSCU le plus élevé. Dans le cas d'*E. coli* K-12, le CAI est utilisé pour prédire l'expressivité des gènes ; il est cependant possible de mesurer n'importe quelle caractéristique biologique, à partir du moment où un biais dans l'usage des codons synonymes des CDS de référence, qui partagent cette caractéristique biologique, est mis en évidence.

Ces calculs d'indices donnent une seule valeur numérique par gène. Graphiquement, les gènes s'organisent donc le long d'un seul axe selon leur valeur d'indice. Comme ils sont sur un seul axe, une seule interprétation est possible : elle résulte de l'équilibre entre deux forces antagonistes, par exemple un usage de codons non optimaux contre un usage de codons optimaux (niveau de traduction élevé). Lors de l'interprétation des résultats, une remarque importante doit être considérée : deux gènes qui présentent un usage des codons similaire auront des valeurs d'indice proche, mais deux gènes qui présentent des valeurs d'indice proches n'ont pas forcément un usage des codons similaires (les usages de différents codons peuvent se compenser). L'analyse devient donc plus fine avec une approche multidimensionnelle : on compare alors deux vecteurs d'utilisation des codons au lieu de deux scalaires. Les résultats deviennent cependant difficiles à visualiser ; c'est pourquoi des méthodes d'analyse multivariée appelées méthodes multifactorielles, sont utilisiées pour faciliter la représentation graphique de données multivariées. Elles réduisent le nombre initial de variables,

[28] D'autres définitions du CBI sont possibles [Morton, 1994]. Le CBI teste alors les fréquences des codons synonymes observées par rapport à l'hypothèse nulle d'un usage des codons synonymes uniforme.

en les remplacant par un petit nombre de composantes synthétiques (facteurs), triées par ordre d'importance. Il existe un second groupe de méthodes d'analyse multivariée, appelées méthodes de classification automatique ou de regroupement, qui permet de réduire le nombre d'individus en formant des groupes homogènes. Ainsi, dans le cas de l'étude du code génétique d'un organisme, on construit différentes classes de gènes sur la base de leur utilisation des codons synonymes. Pour caractériser une typologie ou segmentation des individus, ces méthodes déterminent une ou plusieurs partitions de l'ensemble de départ. Ces méthodes nécessitent de définir un critère de classification (indice de distance entre les individus) et une stratégie d'agrégation (indice d'agglomération, souvent improprement appelé distance, entre les parties ou groupes d'individus). Ce second groupe de méthodes peut être associé au premier, c'est-à-dire qu'une classification peut être réalisée sur les résultats d'une analyse factorielle. Ceci revient, dans un premier temps, à réduire le nombre de variables, et dans un second temps, à réduire le nombre d'individus en les regroupant, l'objectif étant ici de déterminer des jeux de CDS homogènes dans leur usage des codons synonymes.

Le calcul d'indices et l'analyse multivariée ne décrivent que la première phase du travail d'analyse exploratoire de l'usage des codons synonymes. En effet, il ne suffit pas de produire des groupes de gènes ayant un usage des codons synonymes homogènes, encore faut-il être capable de les interpréter, de leur donner un sens biologique en utilisant l'ensemble des informations disponibles pour les gènes d'un groupe particulier. Dans une seconde phase, il est alors possible d'utiliser des méthodes d'analyse bivariée (ou bidimensionnelle), comme la régression linéaire (droites de corrélation) pour étudier la liaison entre deux variables. Deux indices peuvent être comparés, par exemple le CAI, souvent assimilé à l'expressivité des gènes, et l'indice GC_3. On peut aussi comparer l'information portée par un axe factoriel et un indice. Par exemple pour donner un sens biologique à l'information portée par l'axe 1 d'une AFC, on peut étudier sa corrélation avec le CAI ou avec le GC_3.

3.5.2 Méthodes multifactorielles

Les méthodes statistiques *descriptives multidimensionnelles* ont pour objectif de faciliter la représentation graphique de tableaux de données multidimensionnels ; elles utilisent des outils de l'algèbre matricielle [Baccini & Besse, 2002]. Les méthodes multifactorielles cherchent ainsi à réduire la dimension de l'espace de représentation d'un ensemble d'observations, tout en minimisant la perte d'information (Fig. 3.4 p. 126). Elles visent à dégager les tendances les plus fortes d'un nuage de n points à p dimensions, en s'attachant aux relations entre les individus et non pas aux valeurs absolues du tableau. Ces tendances sont représentées par les directions d'allongement maximal du nuage. En pratique, le résultat type de l'analyse d'un nuage de n points à p dimensions est une série de représentations graphiques de ce nuage, sur p axes de projection. Les méthodes multifactorielles comme l'analyse en composantes principale (ACP) ou l'analyse factorielles des correspondances (AFC), diffèrent selon le type de variables considérées (variables quantitatives pour l'ACP, tableau de contingence pour l'AFC), mais permettent toutes de réduire la dimension afin de résumer un tableau np de grande dimension et de révéler ses caractéristiques [Lebart *et al.*, 2000].

Analyse en Composantes Principales (ACP)

L'ACP permet de trouver le meilleur résumé possible pour décrire l'information contenue dans un tableau à n individus et p caractères *numériques*, afin de faciliter l'exploration de ces données multidimensionnelles, selon différents objectifs :

1. réduire efficacement le nombre de dimensions étudiées pour simplifier l'analyse
2. représenter graphiquement les individus en minimisant les déformations du nuage de points
3. représenter graphiquement les variables en explicitant au mieux les liaisons initiales entre ces variables
4. donner des indications sur la nature, la force et la pertinence de ces liaisons

L'ACP consiste à effectuer un changement d'axes dans l'espace des individus qui remplace les variables initiales (corrélées en général) par de nouvelles variables (combinaisons linéaires des variables initiales), non corrélées et de variance maximale : ce sont les *composantes principales*. L'ACP est une méthode factorielle linéaire.

Cette méthode a pour objet la description des données contenues dans un tableau individus-caractères numériques : p caractères sont mesurés sur n individus. Soit e_i, un individu et x_j, un caractère. Chaque individu e_i est alors un point de coordonnées $x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}$ dans l'espace à p dimensions. On cherche le plan de projection orthogonale sur lequel les distances entre deux points seront en moyenne les mieux conservées. Comme l'opération de projection raccourcit toujours les distances, on fixe comme critère de rendre maximale la moyenne des carrés des distances entre les projections. Pour déterminer un nouveau repère de ce plan (*plan principal*), il suffit de trouver deux droites perpendiculaires. Ces droites sont les *axes principaux* du nuage. En projetant e_i qui avait pour coordonnées initiales $x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}$ sur les axes principaux, on obtient de nouvelles coordonnées $c_{i1}, \ldots, c_{ik}, \ldots, c_{ip}$. On construit ainsi de nouveaux caractères $c_1, \ldots, c_k \ldots, c_p$ que l'on appelle les *composantes principales* : chaque composante c_k, qui n'est autre que la liste des coordonnées des n individus sur l'axe k, est une combinaison linéaire des caractères initiaux :

$$c_k = u_{1k}x_1 + \ldots + u_{jk}x_j + \ldots + u_{pk}x_p.$$

Les coefficients $u_{1k}, \ldots, u_{jk}, \ldots, u_{pk}$ forment le kième *facteur principal* u_k [Bouroche & Saporta, 1992]. D'un point de vue mathématique, les principales étapes d'une ACP usuelle sont fondées sur une seconde définition, équivalente à la première, utilisant les outils de l'algèbre matricielle (annexe D p. 399).

Analyse Factorielle des Correspondances (AFC)

L'AFC est une méthode multifactorielle utilisée en particulier pour l'étude de l'usage des codons synonymes de gènes [Grantham *et al.*, 1981, Holm, 1986, Sharp *et al.*, 1988]. Elle permet de calculer les axes de projection les plus aptes à représenter le tableau de données en essayant de minimiser la perte d'information. Cette méthode a été mise au point au début des années 60 à l'université Paris

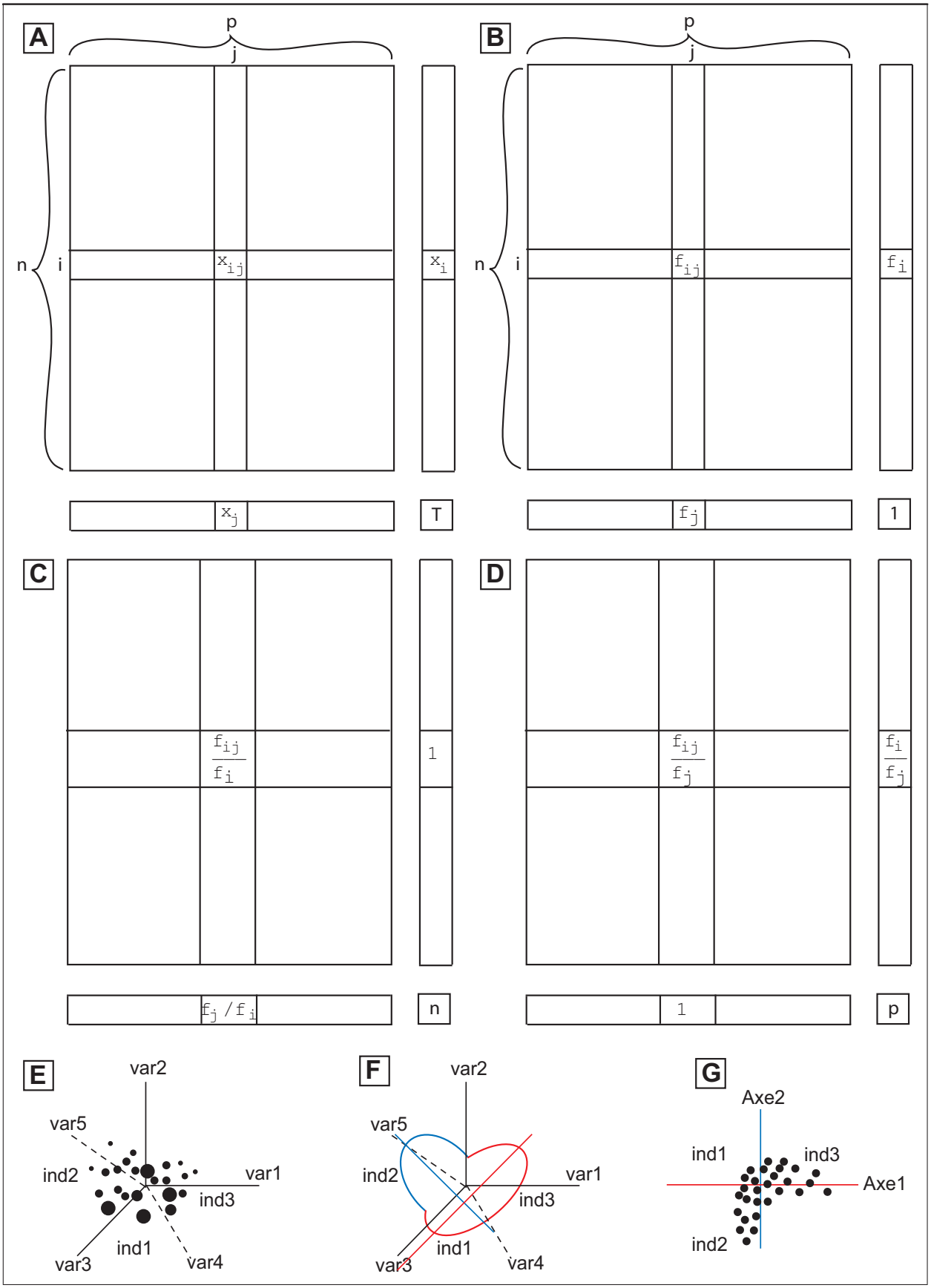


FIG. 3.4 – Principe général de l'AFC (pour la légende voir annexe D p. 399)

A) Tableau de correspondances.

B) Tableau de fréquences conjointes.

C) Profils-lignes.

D) Profils-colonnes.

E) Nuage des n profils-lignes dans l'espace des p variables.

F) Axes de symétrie de l'ellipsoïde d'inertie du nuage d'individus.

G) Projection orthogonal du nuage des individus sur les deux premiers axes factoriels. L'idée est de remplacer les axes définis par les p variables, par les $(p-1)$ axes de symétrie. On projette alors le nuage de points sur les axes de symétrie portant le plus d'inertie. On réduit ainsi la dimension de l'espace de représentation sans modifier le nuage initial.

VI [Benzécri, 1973]. Conçue au départ pour des tableaux de contingence (dans lesquels la case (i, j) contient le nombre d'individus possédant à la fois le caractère i et j), la validité de la méthode s'étend en fait à tout tableau de données vérifiant les conditions suivantes : les données répertoriées dans le tableau sont de même nature et sont toutes positives. Autrement dit, les *marges* (la somme des valeurs d'une ligne x_i ou d'une colonne x_j ; FIG. 3.4 p. 126) doivent avoir une signification. Par exemple, dans un jeu de n CDS décrit par les occurrences des 61 codons (tableau de dimension $n*61$), la somme des valeurs d'une ligne est égale au nombre de codons d'une CDS donnée, et la somme des valeurs d'une colonne est égale au nombre de codons d'un type donné, dans l'ensemble des CDS. L'originalité de l'AFC est de *faire jouer le même rôle aux lignes et aux colonnes* : contrairement à l'ACP, on ne considère pas des individus (lignes) classés selon des variables (colonnes), mais la distribution d'une population répartie selon des caractères interchangeables entre les lignes et les colonnes. Cette conception des données permet d'appliquer la propriété de dualité au tableau de correspondance, et finalement, de représenter les deux populations dans le *même espace*. Ainsi, dans une AFC, on considère successivement les lignes et les colonnes comme *individus* d'une ACP (on ne parlera donc plus d'individus et de variables mais de lignes et de colonnes). L'AFC correspond à une double ACP réalisée sur les profils des lignes et des colonnes selon la métrique du χ^2 (au lieu de la métrique euclidienne dans le cas de l'ACP), afin de révéler les caractéristiques communes entre lignes et/ou entre colonnes. Le χ^2 d'homogénéité constitue le moyen le plus naturel de comparer deux distributions (deux lignes ou deux colonnes) pour déterminer leur degré de ressemblance en fonction de leurs caractères communs. D'un point de vue mathématique, les principales étapes d'une ACP du nuage des n *profils des lignes* dans $\mathbb{R}^p$, réalisée selon la métrique du χ^2, sont résumées en annexe (annexe D p. 399).

AFC de l'usage des codons synonymes (programme *AFCcodons*)

Le programme *AFCcodons* nous a été aimablement fourni par H. Chiapello qui l'a développé au cours de sa thèse [Chiapello, 1999]. Dans le cas de l'étude de l'usage des codons synonymes, les lignes du tableau correspondent aux CDS, et les colonnes aux codons. Le terme de la case (i, j) (ième ligne, jème colonne) contient donc la fréquence relative du codon j dans le gène i (FIG. 3.5 p. 128). Dans le cas du code génétique standard, l'analyse est effectuée sur les 59 codons informatifs, codant 18 acides aminés. Les codons Met et Trp ainsi que les trois codons stop sont exclus. En effet, les acides aminés Met et Trp étant codés par un seul codon (ATG et TGG respectivement), il n'est pas intéressant d'étudier de telles variations. Pour les codons de terminaison, les variations ne doivent pas non plus être étudiées parce que d'un point de vue biologique, ce ne sont pas des codons sens, et que d'un point de vue statistique, cela pose le problème des codons rares (les CDS n'ont qu'un seul codon de terminaison, voir p. 132).

Choix de la fréquence relative des codons synonymes Le choix d'utilisé la fréquence relative des codons ($F_R(abc)$) au lieu de la fréquence absolue ($F(abc)$) présente l'avantage de donner le même poids à chaque gène (TAB. 3.2 p. 89). Par rapport au choix de la fréquence relative des codons, celui

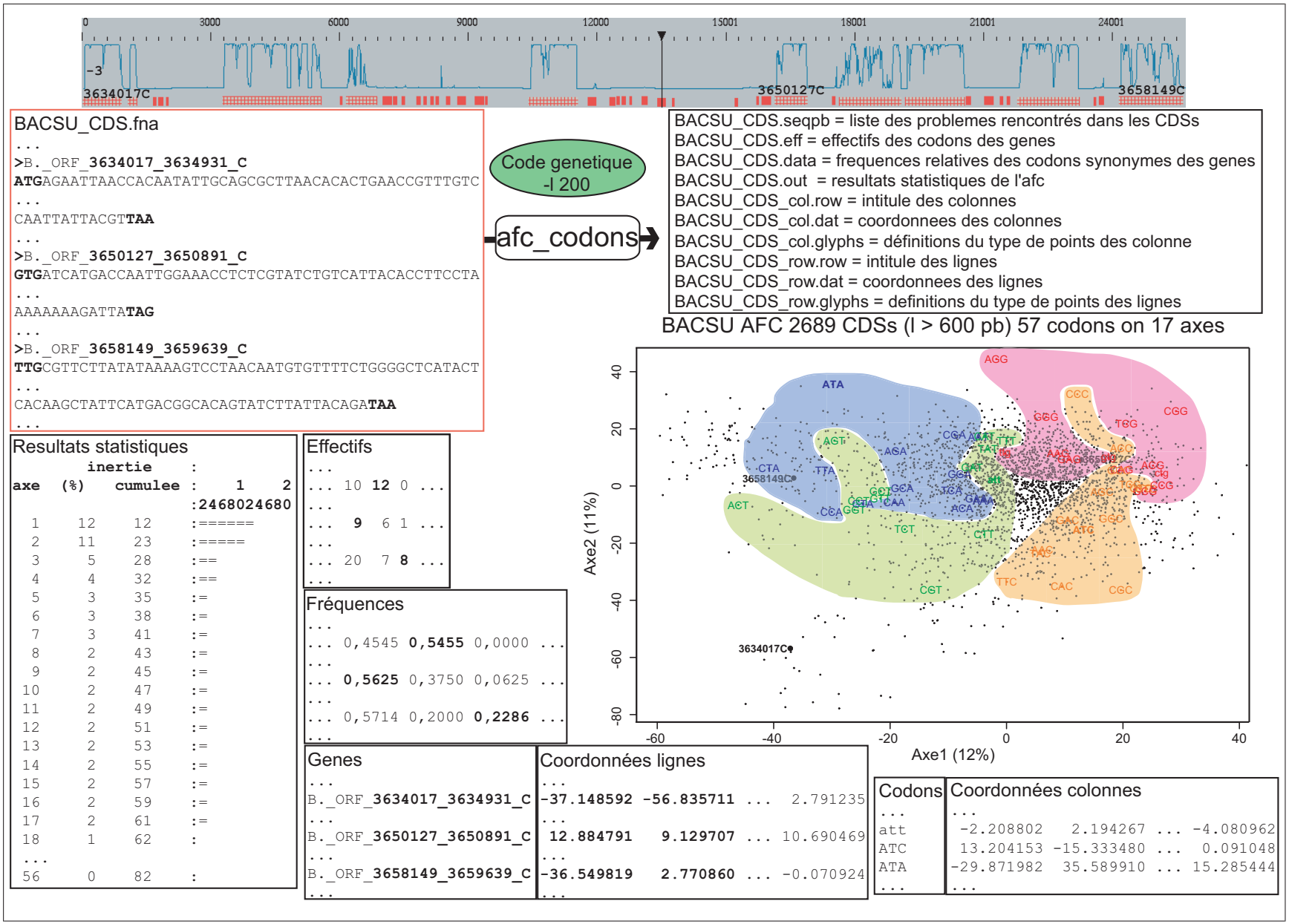


FIG. 3.5 – Le programme *AFCcodons* (pour la légende voir annexe D p. 399)

de la fréquence relative des codons synonymes ($F_{RS}(abc)$) présente l'avantage d'éliminer tout effet lié à la composition en acides aminés dans les séquences traduites des gènes. En effet, la somme des fréquences relatives de tous les codons synonymes codant un même acide aminé est égale à 1. Ainsi, dans le cas du code génétique standard, le total de lignes analysées est de 18 : c'est le nombre d'acides aminés codés par au moins deux codons (la somme des fréquences relatives des codons synonymes codant un même acide aminé dans une CDS vaut 1 ; Tab. 3.2 p. 89). En conséquence, le poids d'un gène dans l'analyse ne dépend ni de sa taille, ni de sa composition en acides aminés, mais seulement de son usage des codons synonymes. On peut donc comparer le biais d'usage des codons des différents gènes de l'échantillon, quelle que soit la composition des protéines correspondantes. Cependant, lorsqu'un acide aminé est rare dans les polypeptides, chacun des codons encodés par cet acide aminé est encore plus rare dans les CDS. Il en résulte que le F_{RS} de ces codons ne prend alors que les valeurs limites 0 ou 1. L'indice F_{RS} est donc sensible aux fluctuations d'échantillonnage (les biais de codons ne sont pas significatifs quand les effectifs sont trop petits). Lors de l'utilisation du programme *AFCcodons* (et d'autres programmes d'analyse descriptive de CDS), il est conseillé d'utiliser un seuil de 300 pb [Karlin, 2001, Garcia-Vallve *et al.*, 2003]. G. Perrière et J. Thioulouse différencient le F_{RS} et le $RSCU$ au niveau de la méthodologie, mais au niveau des résultats, ils ne font plus de différence [Perrière & Thioulouse, 2002]. Il semble équivalent d'utiliser le F_{RS} ou le $RSCU$, lors d'une analyse descriptive de CDS.

Distance entre deux gènes ou entre deux codons Nous avons vu que l'AFC utilise la distance du χ^2 sur le tableau de fréquences *conjointes* de terme courant $f_{ij} = (F_{RS_{ij}}/T$. Dans l'exemple présenté sur la figure 3.5 p. 128, T vaut 45713 soit le produit de 2689 gènes par 17 acides aminés (la cystéine n'a pas été prise en compte ; voir p. 132). Dans le tableau des fréquences *conjointes*, le total général vaut 1 (Fig. 3.4 p. 126). La métrique du χ^2 repose sur la notion de profil. Par exemple, le nuage des profils-lignes (ou points-lignes), noté $N(I)$, est composé des n gènes situés dans l'espace des codons synonymes (59 dimensions), de coordonnée courante f_{ij}/f_i (pondération de la fréquence *conjointe* par la fréquence *marginale* ou poids d'une ligne f_i). Nous cherchons à représenter deux types de distance dans le même espace (sur le même graphique) :

1. la distance $d^2(i, i')$ entre deux gènes i et i' calculée en fonction de leur fréquence d'utilisation des codons synonymes
2. la distance $d^2(j, j')$ entre deux codons j et j' calculée en fonction de leur utilisation dans l'ensemble des gènes de l'échantillon

C'est au cours de cette étape que le poids des lignes (ou des f_i) joue, théoriquement, un rôle dans l'AFC. Mais ici, étant donné que la valeur de ces fréquences f_i sont toutes égales (1/2689), aucune pondération particulière des lignes (gènes) n'est introduite dans le calcul de distances entre deux codons (le calcul du F_{RS} a déjà permis de pondérer les différences entre les longueurs des CDS). La distance entre deux profils-lignes est pondérée par la fréquence *marginale* de chaque codon dans la population des gènes (ou les f_j). Ceci revient à donner une importance comparable aux différents

codons, quelle que soit leur fréquence dans la population des gènes (les variations d'usage des codons peu fréquents ne sont pas masquées par les variations d'usage des codons très fréquents).

Choix de l'AFC Bien que le tableau de données utilisé par le programme *AFCcodons* ne soit pas un tableau de contingence au sens strict, ses valeurs sont toutes positives et de même nature : il est donc possible d'utiliser l'AFC. Cependant, certains chercheurs contestent l'utilisation de l'AFC dans le cadre de l'étude de l'usage des codons synonymes d'un ensemble de gènes [Perrière & Thioulouse, 2002]. G. Perrière préfère utiliser l'AFC sur un tableau de fréquences absolues plutôt que sur un tableau d'usage des codons synonymes (indice *RSCU* ou F_{RS}) car, de son point de vue, les données ne s'y prêtent pas (les marges des lignes sont toutes égales entre elles). En conséquence, l'AFC pondère des lignes qui sont déjà pondérées. Aussi, d'après l'auteur, la métrique du χ^2 perd toute signification lorsqu'elle est utilisée sur des tableaux qui ne sont pas des tableaux d'effectifs [Perrière, 2000]. Il conseille plutôt d'utiliser l'ACP. D'autres auteurs ne font aucune objection quant à l'utilisation de l'AFC sur les *RSCU*[29] [Grantham *et al.*, 1981, Peden, 1999]. Quoi qu'il en soit, utiliser l'ACP ou l'AFC sur un tableau de *RSCU* ou de F_{RS} fournit des résultats similaires, plus faciles à interpréter dans le cas de l'AFC (M.-O. Delorme, communication personnelle). Dans le nuage des points-lignes, chaque point représente un gène, et les gènes dont l'usage des codons est similaire seront voisins sur les axes de projection. L'AFC calcule les axes de projection de telle sorte que la dispersion des points du nuage soit maximale sur le premier plan. En pratique, la coordonnée du gène i sur le kième axe de projection est une combinaison linéaire des fréquences relatives f_{ij} de chaque codon j dans le gène i :
$c_{ik} = (1/\sqrt{\lambda_k}) * \sum_{j=1}^{p}(c_{jk}/fi) * f_{ij} = (1/\sqrt{\lambda_k} * (a_1 f_{i1} + a_2 f_{i2} + \ldots + a_{59} f_{i59})$
où λ_k et a_j sont des coefficients calculés par l'AFC et $p \leq 59$. Ainsi, l'AFC présente l'avantage de simplifier l'interprétation des proximités entre points-lignes et points-colonnes du graphique : au sein des nuages N(I) et N(J), la proximité entre la projection d'un point-ligne (un gène) et d'un point-colonne (un codon) assure que le codon joue un rôle important pour le gène correspondant, et vice-versa. Cette observation n'est cependant plus vraie lorsque l'on se rapproche de l'origine du plan de projection : la proximité d'un point-ligne et d'un point-colonne peut être alors totalement fortuite, et ne correspondre à aucune proximité réelle dans l'espace.

Interprétation des résultats L'interprétation des résultats de ce type de méthode est délicate ; aussi nécessite-t-elle une certaine méthodologie. Dans une première étape, il est nécessaire d'apprécier la qualité et l'importance des résultats. On s'intéresse tout d'abord à la mesure globale du pourcentage d'inertie porté par chacun des axes de projection. Le programme *AFCcodons* ne retient que les axes portant un pourcentage d'inertie significatif, c'est-à-dire supérieur ou égal à 2%. Ce seuil est défini selon le critère empirique de la règle de Kaiser (le pourcentage d'inertie doit être supérieur à $100/p$ d'où $100/59 = 1,7\%$ [Baccini & Besse, 2002]). Par ailleurs, parmi les axes portant plus de 2% d'inertie, seuls ceux permettant de révéler une tendance générale dans le

[29] Dans le cas de l'AFC sur les *RSCU*, la pondération des lignes ne sert à rien, mais elle ne fausse pas les résultats.

nuage de points seront utilisés pour la projection des nuages de points. Dans cet objectif, le critère empirique de la *règle de l'éboulis des valeurs propres* (ou *règle du coude*) est utilisé : tant que les écarts entre les inerties portées par deux axes consécutifs sont importants, des tendances générales sont mises en évidence. L'examen des écarts prend fin lorsqu'un écart entre deux axes successifs devient négligeable [Baccini & Besse, 2002]. Dans l'exemple de la figure 3.5 p. 128, les deux, voire les quatre premiers axes de projection seront examinés. On s'intéresse ensuite à la représentation du nuage de points (gènes) sur les axes conservés. Par exemple, si les quatre premiers axes sont retenus, une combinaison de six représentations à deux dimensions est réalisée. Deux gènes qui présentent un usage des codons synonymes similaire seront voisins sur le graphe. Le contraire n'est pas nécessairement vrai, car la représentation plane sur seulement deux axes (souvent les deux premiers axes de l'AFC) masque les autres orientations du nuage qui ont, généralement, une signification moins forte. Un défaut de perspective peut rendre proches deux points qui seraient très éloignés suivant un axe d'inertie inférieure. Parmi les indices d'aide à l'interprétation de l'AFC, le calcul des *contributions* des points-lignes et colonnes, à l'inertie expliquée par les axes, est une mesure très importante qui permet de classer les points selon le rôle plus ou moins grand qu'ils ont joué dans la formation d'un axe. On s'intéresse enfin à la représentation superposée des gènes et des codons sur les axes de projection conservés. La représentation de l'espace dual permet de corréler une tendance observée dans l'usage des codons synonymes avec les codons préférentiellement utilisés par les gènes correspondants.

La seconde phase de l'interprétation est externe au programme *AFCcodons*, puisqu'il s'agit de chercher à comprendre les tendances majeures révélées par l'AFC, tant du point de vue des gènes que des codons. C'est souvent la partie la plus fastidieuse du travail car elle nécessite d'analyser finement les différences d'usage des codons révélées par l'AFC, et de comprendre s'il existe un lien avec des fonctions et/ou des caractéristiques biologiques. En général, des variables et/ou des individus supplémentaires sont superposés dans la représentation graphique. Par exemple, l'ajout de la variable *précoce-tardif* (CDS), qui consiste à attribuer la valeur -1 à une CDS du brin tardif et $+1$ à une CDS du brin précoce, permet de les colorer, sur la représentation graphique de l'AFC, en rouge si elles sont sur le brin précoce et en bleu sur le brin tardif. Il devient alors possible de répondre à la question : « Est ce que l'un des deux premiers axes de l'AFC révèle un biais de codons lié à la réplication ? » De même, si l'on possède des jeux de CDS hautement exprimés, xénologues, essentiels, etc., on pourra visualiser s'il existe un lien entre ces caractéristiques biologiques et les tendances révélées par les axes informatifs de l'AFC. Comme nous l'avons vu plus haut, pour établir une liaison entre deux variables de manière plus précise, il est possible de représenter la valeur de l'axe 1 des gènes contre celle d'un indice mesuré (GC_3, CAI), afin de déterminer si l'inertie portée par l'axe est corrélée à cet indice. Enfin l'observation des différents nuages de points permet de déterminer, en première approximation, le nombre de groupes de gènes qui semblent posséder des biais communs dans l'usage des codons synonymes ; ce nombre, k, peut alors être utilisé par un programme de partitionnement automatique (voir p. 133).

Une dernière remarque concerne l'utilisation du terme *biais d'usage des codons*. Plusieurs au-

teurs utilisent ce terme pour caractériser un usage des codons qui s'éloigne le plus d'un usage où chaque codon synonyme est équiprobable (les indices *RSCU* valent 1 pour tous les codons). Au cours de l'analyse des résultats de l'AFC, un gène est considéré comme ayant un usage des codons biaisé si les fréquences de ses codons synonymes s'écartent fortement des fréquences moyennes des codons synonymes observés dans les autres gènes de l'échantillon. Il s'agit donc des gènes s'écartant significativement de l'origine des différents plans de projection de l'AFC.

Avantage et inconvénient de la méthode L'AFC est une technique puissante permettant, entre autres, d'analyser l'usage des codons d'un ensemble de gènes. Elle est particulièrement bien adaptée dans le cadre de l'analyse de grands jeux de données tels que l'ensemble des gènes annotés d'un génome complet. A la différence d'un calcul d'indice comme le CAI, l'AFC ne nécessite aucune connaissance *a priori* (pas de phase d'apprentissage) et permet de révéler *plusieurs* tendances (une sur chaque dimension informative). Etudier l'usage des codons synonymes est une première étape, incontournable, à la caractérisation de groupes de gènes qui peuvent avoir en commun une histoire, une localisation, un niveau d'expression, etc. Bien sûr, relier un usage des codons synonymes particulier à une explication biologique est loin d'être une tâche facile ; elle n'en reste pas moins très intéressante. Dans le cadre de l'étude de l'usage des codons synonymes, cette méthode possède essentiellement trois inconvénients.

1. L'analyse étant réalisée gène par gène, il n'est pas possible de l'utiliser sur des gènes contenant un nombre insuffisant de codons. Dans le cadre d'une analyse *globale* de l'usage des codons d'un génome, l'utilisation d'un jeu de CDS incomplet ne modifiera pas les tendances générales des nuages.
2. Le deuxième inconvénient résulte du problème des acides aminés rares, dont que la cystéine. En effet, beaucoup de gènes contiennent très peu de codons cystéine (moins de trois) et les fluctuations d'échantillonnage faussent alors l'interprétation des résultats. La représentation graphique de l'AFC peut révéler trois groupes de gènes : un groupe dont les gènes utilisent dans 100% des cas le codon TGC, un groupe dont les gènes utilisent TGC et TGT ou ne contiennent pas de cystéine, et un groupe dont les gènes utilisent dans 100% des cas le codon TGT [Perrière & Thioulouse, 2002]. Cette valeur de 100% n'a cependant pas de sens puisqu'il n'y a qu'un seul codon cystéine dans les gènes concernés. Afin de s'affranchir de ce biais artificiel, on supprime généralement, de l'analyse réalisée avec l'AFC, les codons des acides aminés rares, par exemple les deux codons TGC et TGT de la cystéine [Moszer, 1996].
3. Le troisième inconvénient, [Greenacre, 1984, Benzécri, 1973], est due au problème de l'effet Guttman. Sur chaque axe d'une analyse factorielle, les individus sont ordonnés selon les valeurs de la tendance révélée et l'on s'attend, en général, à voir apparaître un gradient de points (ellipsoïde). Il arrive cependant que le nuage projeté sur les deux premiers axes ait une forme de fer à cheval (parabolique), qui suggère que les axes 1 et 2 ne soient pas complètement indépendants (Fig. p. 218). Cette interprétation semble contradictoire avec les fondements

de l'AFC où les axes factoriels sont définis comme de nouvelles variables, *non corrélées entre elles* et donc indépendantes. Deux explications à ce phénomène sont possibles :

(a) Sur certaines données, l'AFC produit une relation artefactuelle entre les deux premiers facteurs.

(b) Sur d'autres données, il existe réellement une relation entre les deux axes, qu'il faut chercher à expliquer.

Dans le cas d'un artefact, l'effet Guttman peut traduire, par exemple, une certaine redondance entre les coordonnées de l'axe 1 et celles de l'axe2. Le deuxième facteur pourrait être une fonction du second degré du premier facteur, le troisième facteur pourrait être une fonction du troisième degré du second facteur, etc. L'examen du second facteur affine cependant l'interprétation du premier axe : le premier axe oppose les valeurs extrêmes et le deuxième oppose les valeurs intermédiaires aux valeurs extrêmes [Lebart *et al.*, 2000].

Dans le cas d'une relation réelle entre les deux axes, si l'inertie de l'axe 1 s'explique, par exemple, par la composition des gènes (A+T riches et G+C riches), et que celle de l'axe 2 s'explique par leur expressivité (expression forte, constitutive, etc.), on peut alors imaginer qu'il y ait une relation entre la composition des gènes et leur niveau d'expression (voir p. 216).

3.5.3 Méthodes de classification automatique

L'objectif des méthodes de classification automatique (*clustering analysis* (*CA*)) consiste à construire une partition ou une suite de partitions emboîtées (hiérarchie) d'un ensemble de n objets caractérisés par un certain nombre de mesures. Il s'agit de caractériser les sous-ensembles d'objets qui soient à la fois homogènes (c'est-à-dire qu'il faut minimiser l'inertie intra-classe) et bien séparés (c'est-à-dire qu'il faut maximiser la distance inter-classe) : les éléments appartenant à un sous-ensemble se ressemblent, et deux éléments appartenant à des sous-ensembles distincts sont différents. Nous présenterons ici les méthodes de classification les plus classiques, à savoir la classification hiérarchique ascendante et le partitionnement ; il existe cependant de nombreuses autres méthodes de classification.

Principe et exemples d'application

Dans un premier temps, des données de proximités entre individus doivent être générées à partir des données brutes (n individus et p variables) et d'un critère de proximité. Dans un second temps, la méthode construit des partitions ou des suites de partitions emboîtées des n objets (hiérarchies) dont une seule sera conservée. La démarche générale d'une classification automatique consiste à effectuer les opérations suivantes :

1. Sélectionner un échantillon de n individus parmi lesquels on cherche à construire des groupes, et effectuer p mesures sur chaque individu afin d'obtenir un tableau rectangulaire de valeurs numériques (données brutes). Ces dernières peuvent être les mesures d'une ou plusieurs variables observées (tableaux numériques de variables quantitatives comme la taille ou le poids,

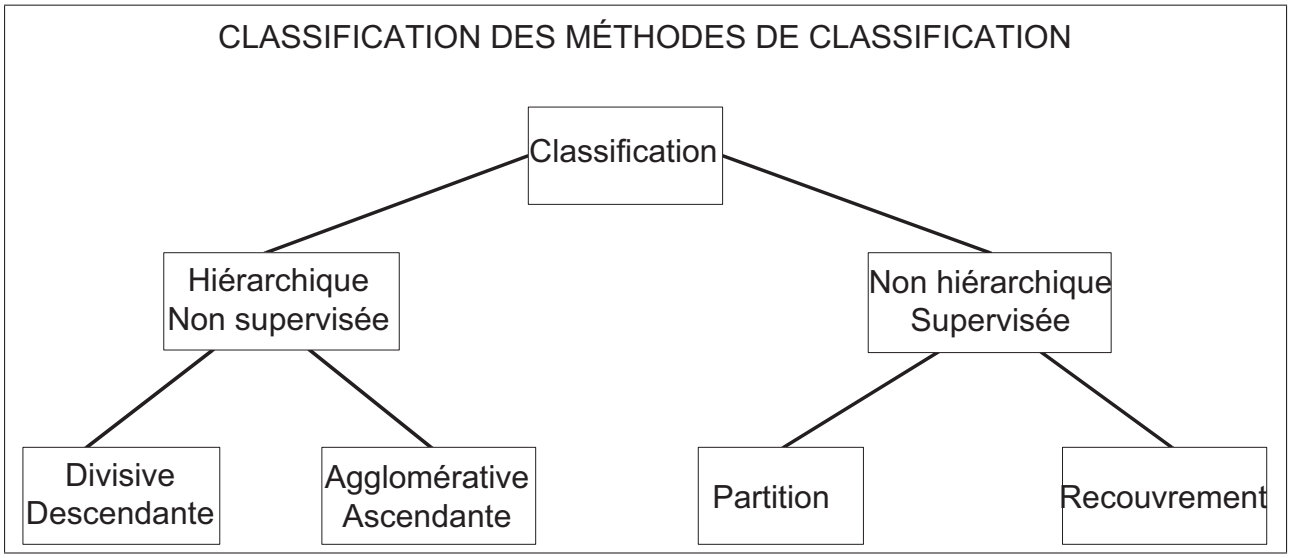


FIG. 3.6 – Classification des méthodes de classification

Il existe deux types de classifications : la *classification hiérarchique*, au sein de laquelle les groupes créés sont emboîtés les uns dans les autres de façon hiérarchique (classification des êtres vivants), et la *classification non hiérarchique*, qui ne cherche pas à structurer les sous-ensembles entre eux de manière hiérarchique.

En classification hiérarchique, deux types de méthodes peuvent être mises en œuvre :

(1) une méthode descendante (divisive) qui part de l'ensemble de tous les éléments pour les fractionner en un certain nombre de sous-ensembles, ces derniers étant alors eux-mêmes fractionnés en sous-ensemble plus petits. Le processus est itéré jusqu'à l'obtention d'éléments individuels (Règne → Embranchement → Classe → ... → Genre → Espèce)

(2) une méthode ascendante (agglomérative) qui part des éléments individuels pour les regrouper en sous-ensembles, avant d'appliquer récursivement un algorithme de regroupement des sous-ensembles jusqu'à obtention d'un seul ensemble contenant tous les éléments (Espèce → Genre → Classe → ... → Embranchement → Règne)

Une hiérarchie est ainsi le plus souvent construite de manière itérative, en regroupant deux à deux les objets (individus ou groupes d'individus) qui se ressemblent le plus.

En classification non hiérarchique, il est possible de considérer que chaque élément ne fait partie que d'un sous-ensemble (on parle alors de partition), ou au contraire, qu'il peut faire partie de plusieurs sous-ensembles (on parle alors de classification recouvrante) ; dans ce cas, une probabilité d'appartenance à chaque groupe est calculée.

Les méthodes de classification non hiérarchique sont qualifiées de supervisées car l'utilisateur doit préciser le nombre de classes *a priori*, alors que les méthodes de classification hiérarchiques sont non supervisées car l'utilisateur ne donne pas d'information sur le nombre de classes. Une dernière remarque concerne la différence entre les mots classement, classification et regroupement : classification est parfois employé dans le sens de classement : une méthode de classement est une façon d'associer à chaque individu une classe prédéterminée (codant vs non-codant). Le mot regroupement (*clustering*) correspond à la démarche de déterminer des classes à partir des individus (classifier en deux groupes dont on ne sait pas à priori ce qu'ils représentent). Cette figure a été extraite de `ina.eivd.ch/collaborateurs/etd/presentations.dir/tr_classification.pdf`

tableaux de contingence comme les fréquences de codons dans les gènes, tableau de présence-absence, etc.), ou les coordonnées des individus sur les axes d'une analyse factorielle (ACP, AFC).

2. Choisir un critère de classification, c'est-à-dire définir une proximité entre individus : une distance (ce qui respecte l'inégalité triangulaire) ou un indice de distance (ce qui ne respecte pas forcément l'inégalité triangulaire). La proximité peut être mesurée par :
   - Une distance $d(ij)$ telle que $d(ij) = d(ji)$; $d(ij) \geq 0$; $d(ij) = 0 \leftrightarrow i = j$; $d(ij) \leq d(ik) + d(kj)$. Pour les tableaux numériques, la distance euclidienne est utilisée. Pour les tableaux de comptage, la distance du χ^2 est préférée. Cette distance introduit une division par les effectifs marginaux et présente donc l'avantage de comparer des profils.
   - Une dissimilitude $ds(ij)$ telle que $ds(ij) = ds(ji)$; $ds(ij) \geq 0$; $ds(ii) = 0$. Pour les données d'alignement de séquences, on dénombre les nucléotides qui diffèrent entre deux séquences. L'indice de dissimilitude donne une estimation du le nombre de substitutions nucléotidiques par site. Il existe plusieurs modèles de substitutions nucléotidiques tels que celui de Jukes et Cantor ou encore celui de Kimura [Nei & Kumar, 2000a].
   - Une similitude $s(ij)$ telle que $s(ij) = s(ji)$; $s(ij) \geq 0$; $s(ii) \geq s(ij)$. Pour les données présence-absence, on utilise par exemple l'indice de Jaccard [Croquette & Carlier, 1999].

   Certaines méthodes de classification (hiérarchique ascendante) nécessitent des données de proximité entre individus d'un même ensemble, qui sont contenues dans des matrices carrées $(n * n)$ et symétriques, dont les termes courants appartiennent à l'ensemble des réels.

3. Choisir un type de classification (hiérarchique, partition) et, éventuellement, des contraintes additionnelles (nombre de groupes à former, nombre maximum d'éléments dans un groupe, etc.). *A priori*, pour trouver la meilleure partition, il suffit d'estimer la qualité de toutes les partitions possibles et de choisir celle qui optimise le critère de qualité (homogénéité des groupes). Cependant, l'approche exacte est confrontée aux problèmes de combinatoires. Le nombre de partitions possibles d'un ensemble de n éléments croît exponentiellement avec n. Pour contourner ce problème, il est nécessaire d'utiliser des heuristiques. Il existe deux structures principales de classification : hiérarchique et non hiérarchique. Pour les méthodes hiérarchiques, le nombre k de groupes n'est pas déterminé *a priori* : on dit qu'elles sont non supervisées (Fig. 3.6 p. 134). Ces méthodes ont une complexité de calcul en n^3 et ne remettent pas en cause la qualité des partitions construites à chaque étape. En classification non hiérarchique, le nombre k de groupes est déterminé *a priori*. Ces méthodes ont une complexité de calcul en n et estiment la qualité des partitions à chaque itération.

4. Choisir un critère global de mesure de qualité d'une classification, basé sur l'éloignement entre les groupes ou l'homogénéité des individus à l'intérieur des groupes. En fonction des données de proximité et du type de classification, le critère de mesure de qualité est différent. Par exemple, dans un espace euclidien, l'inertie (la dispersion globale du nuage de points) est mesurée. La relation de Huyghens décrit l'inertie totale comme la somme de l'inertie intra-

groupe et de l'inertie inter-groupe. Un critère global de mesure de qualité d'une classification peut consister à chercher la classification qui minimise l'inertie intra-groupe (compacité des groupes), ce qui équivaut à maximiser l'inertie inter-groupe (éloignement entre les groupes) [Croquette & Carlier, 1999].

Les méthodes de classification automatique sont utilisées dans le cadre de l'analyse multivariée. En effet, un ensemble d'individus décrit par deux variables peut être représenté dans un plan ; si des groupes apparaissent distinctement, une classification manuelle établie par l'utilisateur est amplement suffisante. En revanche, lorsque l'on s'intéresse aux coordonnées calculées par une AFC dans le cadre de l'analyse de l'usage des codons synonymes d'un ensemble de gènes, l'utilisation d'une méthode de classification automatique est recommandée. La détermination précise des différents groupes de gènes, représentés par la projection du nuage de points sur les plans factoriels, est alors une opération délicate. Nous avons vu que les représentations euclidiennes d'un tableau de correspondances nécessitent souvent un espace pourvu de plus de trois dimensions, ce qui complique la recherche de voisinage entre objets. Puisqu'il est souvent difficile, à l'œil nu, de délimiter les groupes de manière objective, on comprend l'intérêt d'utiliser une méthode de classification qui va regrouper automatiquement les points proches dans l'espace. Les gènes sont alors classés en fonction de leurs coordonnées de projection sur les différents axes factoriels et non pas en fonction de la distance entre deux gènes. Nous présenterons rapidement les méthodes usuelles de la classification hiérarchique, puis nous nous concentrerons sur les méthodes de partitionnement par réallocation dynamique. En effet, du fait de leur complexité de calcul, les méthodes de partitionnement sont plus adaptées que les méthodes de classification hiérarchique pour traiter les grands échantillons (n gènes à p variables dans le cas de l'étude de l'usage des codons).

Classification hiérarchique ascendante

A toute hiérarchie correspond un arbre de classification qui peut être vu comme une suite de partitions emboîtées. Nous considérons ici des arbres dichotomiques, c'est-à-dire qu'une branche ne peut se diviser qu'en deux branches filles. Une feuille correspond au nœud terminal d'une branche (une séquence observée). Un arbre peut être enraciné (avoir un ancêtre commun à toutes les séquences observées) ou non. Les fondements de la hiérarchie ascendante reposent sur une stratégie d'agrégation. Cette stratégie définit, entre autres, un indice d'agglomération entre parties ou groupes d'individus, improprement appelé *distance*, car ces indices ne respectent pas forcément l'inégalité triangulaire (annexe E p. 409). Un indice d'agglomération est calculé à partir d'une distance entre deux individus (ou d'un indice de distance). Cette méthode met en œuvre un algorithme itératif à deux étapes :

1. on cherche tout d'abord à constituer un groupe par agglomération de deux éléments (individu ou groupe d'individus)
2. puis on met à jour le tableau d'indices d'agglomération en calculant l'indice d'agglomération entre le nouveau groupe constitué et tous les autres éléments pris individuellement (individu

ou groupe)

A chaque itération, on se sert de l'indice d'agglomération (*distance*) entre deux groupes afin d'homogénéiser un nouveau groupe (on agrège les deux groupes qui ont l'indice d'agglomération le plus faible). Ce processus d'agrégation pas à pas est basé sur un critère d'homogénéité à l'intérieur d'un groupe : c'est pourquoi il est inutile de mesurer la qualité globale de la nouvelle partition à chaque itération. La méthode UPGMA (*Unweighted Pair-Group Method using arithmetic Averages* [Sneath & Sokal, 1973, Nei & Kumar, 2000c]) consiste à agréger les deux éléments les plus proches selon l'indice d'agglomération du saut moyen (*average linkage* est la *distance* moyenne entre deux groupes,voir annexe E p. 409).

Cette méthode utilisée pour la reconstruction d'arbre phylogénétique a pour inconvénient majeur de reposer sur l'hypothèse que les vitesses d'évolution sont les mêmes sur les différentes branches de l'arbre, ce qui peut conduire à une topologie incorrecte de l'arbre reconstruit. La méthode NJ développée par N. Saitou et M. Nei (*Neighbor-Joining*, version simplifiée de la méthode de l'évolution minimum [Nei & Kumar, 2000c]), tente de corriger la méthode UPGMA afin d'autoriser un taux de mutation différent sur les branches de l'arbre. Dans cette méthode, les séquences sont représentées au départ par un arbre en étoile que l'on cherche à étirer, et non pas par des points que l'on cherche à encercler. L'indice d'agglomération n'est pas calculé à l'aide d'un indice classique de distance entre deux groupes de séquences (saut moyen, saut maximum, saut minimum ; voir annexe E p. 409) puisqu'il repose sur la distance entre deux feuilles de l'arbre, les longueurs des branches étant additives. L'indice d'agglomération[30] est la distance entre deux feuilles, corrigée afin de prendre en compte la divergence moyenne de chacune des séquences avec les autres. Le concept important de cette méthode est celui des *voisins* qui sont définis comme des groupes connectés par un nœud dans un arbre non enraciné. Ce concept permet de tenir compte du fait que les deux feuilles les plus proches, selon leur indice de dissimilitude, ne sont pas forcément *voisines* sur l'arbre [Durbin *et al.*, 2001c]. A chaque itération, une matrice des distances d_{ij} entre les feuilles d'un arbre est calculée, à partir de laquelle est déduite une matrice d'agglomération D_{ij}. On agrège les deux feuilles *voisines* les plus proches selon l'indice d'agglomération le plus faible. Le nouveau groupe est représenté par une feuille k, précédemment nœud parent des deux feuilles agrégées i, j. Autrement dit, lorsque deux feuilles sont regroupées, le nœud représentant leur ancêtre commun est ajouté (au départ il n'y a qu'un nœud au centre de l'étoile) et les deux feuilles sont enlevées. Ce processus convertit l'ancêtre commun en une feuille dans un arbre dont la taille a diminué ($d_{km} = 1/2(d_{im} + d_{jm} - d_{ij})$ avec m une autre feuille). Cette méthode nécessite le taxon d'un extra-groupe enraciner l'arbre (Fig. 1.2 p. 53).

Les méthodes de classification hiérarchique sont gourmandes en ressources informatiques. C'est pourquoi les méthodes UPGMA et NJ sont appliquées à la classification de gènes dans le cadre d'études de phylogénie moléculaire, plutôt que dans celui de l'analyse de l'usage des codons. Toutes

[30] En pratique, on calcule l'indice d'agglomération D_{ij} comme la distance entre deux feuilles d_{ij} corrigée par la divergence moyenne de la première feuille r_i (moyenne des distances entre i et les autres feuilles k prises individuellement) et de la seconde : $D_{ij} = d_{ij} - (r_i + r_j)$ avec $r_i = (1/L - 2) \sum_{k \in L} d_{ik}$ et L le nombre de feuilles.

les méthodes de reconstruction d'arbre phylogénétique font l'hypothèse que les sites des séquences évoluent indépendamment ; cependant on peut distinguer deux approches principales : celles qui reposent sur les distances entre les séquences prises deux à deux (UPGMA et NJ) et celles qui n'utilisent pas les distances (parcimonie, maximum de vraisemblance [Nei & Kumar, 2000b]). Les étapes classiques de l'approche fondée sur les distances sont constituées d'un alignement multiple des séquences, d'un calcul de matrices de dissimilitudes entre tous les couples de séquences, et d'une classification hiérarchique ascendante.

Partitionnement

Les méthodes de partitionnement permettent de traiter rapidement des ensembles d'effectifs élevés, en optimisant localement un critère tel que l'inertie intra-classe[31]. On cherche généralement à minimiser l'inertie intra-classe, le nombre k de groupes étant fixé car le maximum d'inertie inter-classe est obtenu pour la partition ayant comme classes les singletons. L'inertie intra-classe ne permet pas de comparer deux partitions ayant des nombres de classes différents. Le but des méthodes de partitionnement est de construire une partition unique des objets en k groupes, k étant fixé *a priori*, ou bien déterminé par la méthode. L'idée centrale est de choisir une partition initiale des objets et de déplacer les objets d'un groupe à l'autre pour obtenir une partition stable. Plusieurs algorithmes de partitionnement ont été développés, qui diffèrent dans le choix de la partition initiale, dans la définition de la meilleure partition, et dans la méthode utilisée pour améliorer la partition.

Méthode des centres mobiles Différents types d'algorithmes ont été définis autour du principe de réallocation dynamique des individus à k centres de classes (le nombre de classes k étant fixé *a priori*), les centres étant recalculés à chaque itération. Le principe général de la méthode des centres mobiles peut-être résumé comme suit : dans l'espace $\mathbb{R}^p$ muni de la métrique identité (métrique euclidienne classique), on définit le nuage de n individus N(I) par un tableau de données de taille $n * p$ où p est le nombre de variables.

1. Initialisation :
   (a) choix d'une configuration initiale de k centres de groupes *ou*
   (b) choix d'une partition initiale d'objets en k groupes et calcul de leur centre
2. Itération :
   (a) allocation des objets au groupe dont le centre est le plus proche. On obtient alors une partition en k groupes (ou moins si l'un des groupes est vide) *et*
   (b) calcul des nouveaux centres des groupes (si un groupe s'est vidé on peut choisir un centre supplémentaire). On dit qu'un groupe est vide quand aucun point n'a été attribué à un centre
3. Condition d'arrêt : la convergence est atteinte quand

[31] Les individus sont des points de l'espace euclidien $\mathbb{R}^p$.

(a) deux étapes successives ne modifient pas les classes et les centres

(b) la valeur du critère global de qualité de la partition (l'inertie inter- ou intra- classe) ne varie plus significativement entre deux itérations *ou*

(c) le nombre d'itérations fixé est atteint

Partant de ce principe général de la méthode des centres mobiles, plusieurs variantes de l'algorithme de partitionnement ont été développées.

Variantes de la méthode des centres mobiles Pour l'algorithme de Forgy [Forgy, 1965], l'initialisation est réalisée à partir de k individus de l'ensemble qui représentent les k centres de gravité initiaux. Ces k individus sont soit tirés au hasard, soit déterminés par un expert. Au cours du processus itératif, on choisit d'affecter un point au groupe pour lequel la distance entre le point et le centre de gravité du groupe est minimale. La condition d'arrêt est la stabilisation de la partition ou la minimisation du critère d'inertie intra-classe (somme des k inerties à l'intérieur d'un groupe).

Dans le cas de l'algorithme des *K-means*, il existe deux variantes. La configuration initiale de MacQueen [MacQueen, 1967] consiste à choisir les k premiers objets de l'échantillon. Au cours du processus itératif, au lieu d'attendre que tous les individus aient été affectés à une classe pour recalculer les centres de gravité, ces derniers sont recalculés à chaque nouvelle affectation d'un point. L'algorithme s'arrête lorsque le nombre d'itérations fixé, ou le maximum du critère d'inertie inter-classe (somme des k inerties entre les centres de gravité des groupes) est atteint. La variante d'Hartigan [Hartigan & Wong, 1979] ressemble à l'algorithme de Forgy à la différence qu'un nombre d'itérations maximum est fixé (Fig. 3.7 p. 140).

La méthode des nuées dynamiques (*dynamic clusters*) peut être considérée comme une généralisation de la méthode des centres mobiles ; elle admet donc comme cas particulier la méthode de Forgy. Au lieu de représenter un groupe par son centre de gravité, on le représente par Q individus (les plus centraux) formant un *noyau* (*kernel*). Chacun des K groupes est donc représenté par Q individus appelés étalons. Soient E l'ensemble des objets à classer et E_k les étalons de la kième classe. Nous allons décrire l'algorithme tel qu'il est présenté dans l'article original d'E. Diday [Diday, 1971, Croquette & Carlier, 1999] :

1. Initialisation : pour chaque $k \in \{1, 2, \ldots, K\}$, tirer au hasard Q étalons (ou les choisir expressément), et définir ainsi $L^{(0)} = \{E_1^{(0)}, E_2^{(0)}, \ldots, E_K^{(0)}\}$.
2. Itération : connaissant $L^{(r)}$ calculer $L^{(r+1)}$.
   - Pour chaque $k \in \{1, 2, \ldots, K\}$ et pour chaque $x \in E$ calculer $D(x, E_k^{(r)})$.
   - Pour chaque $k \in \{1, 2, \ldots, K\}$ et pour chaque $x \in E$ minimiser $D(x, E_k^{(r)})$, et faire ainsi la partition de E en K classes $C_k^{(r)}$.
   - Pour chaque $k \in \{1, 2, \ldots, K\}$ et pour chaque $x \in E$ choisir les Q_k objets de E qui minimisent $R(x, k, L^{(r)})$, et définir ainsi

$$L^{(r+1)} = \{E_1^{(r+1)}, E_2^{(r+1)}, \ldots, E_K^{(r+1)}\}.$$

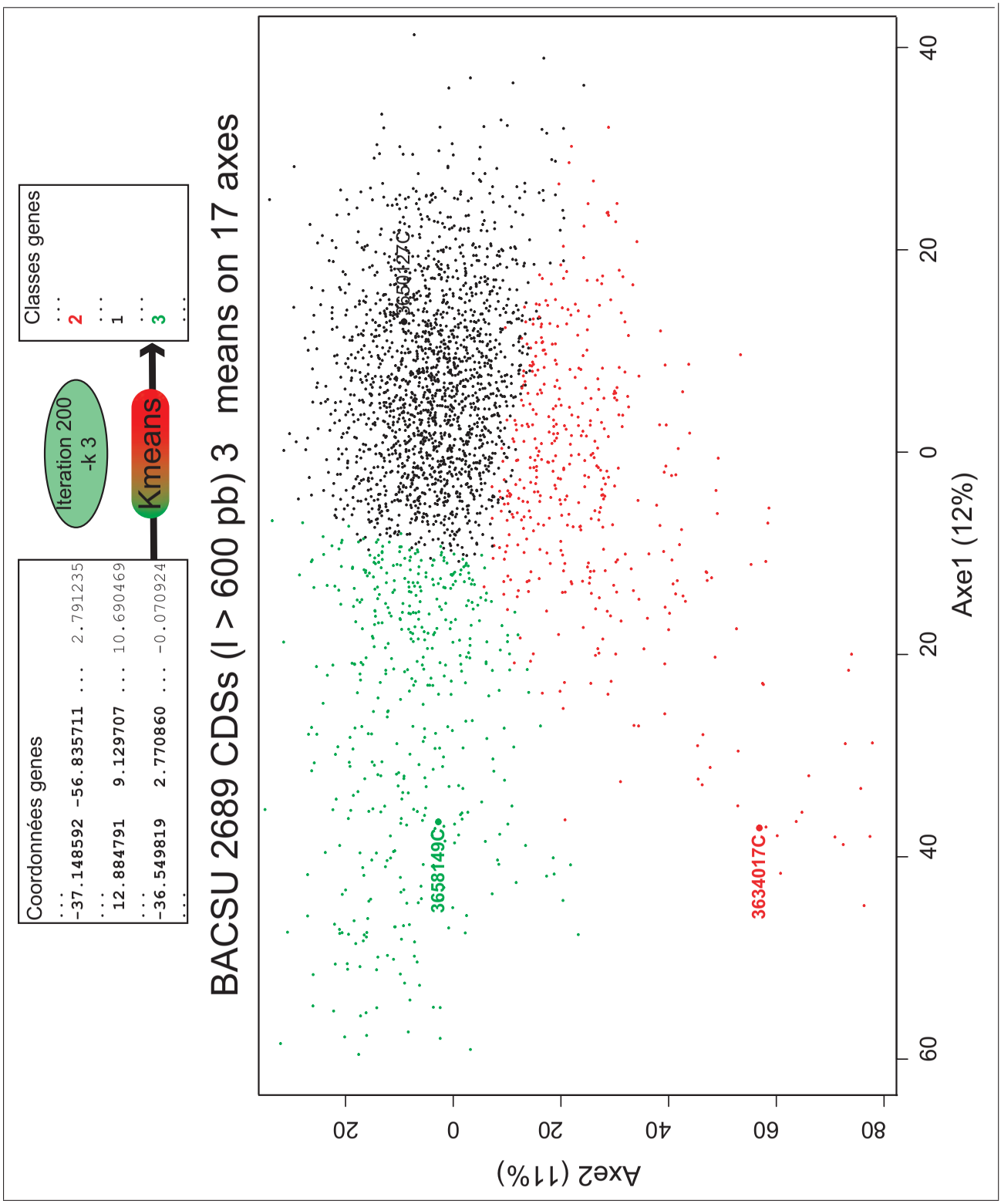


FIG. 3.7 – Partitionnement automatique par la variante du K-means des centres mobiles annexe E

3. Condition d'arrêt : $L^{(r)} = L^{(r+1)}$.

On peut choisir de calculer la distance d'un individu à un noyau de la manière suivante : $D(x, E) = \sum_{y \in E} d(x, y)$. De même, on peut choisir la fonction d'agrégation-écartement comme :

$$\begin{aligned} R(x, k, L) &= D(x, C_k) \text{ ou} \\ R(x, k, L) &= \frac{D(x, E_k) D(x, C_k)}{[\sum_{i=1}^{K} D(x, E_i)]^2}. \end{aligned}$$

Dans le dernier cas, $D(x, E_k)$, $D(x, C_k)$ et $\sum_{i=1}^{K} D(x, E_i)$ auront pour effet respectivement d'agréger les étalons, de ramener les étalons vers le centre de leur classe et d'écarter les E_i entre eux. On dira que $x \in C_k$ est un élément améliorant de E_k dans E s'il existe $q \in \{1, 2, \ldots, Q_k\}$ tel que : $R(x, k, L) < R(e_{qk}, k, L)$ ($e_{qk} \in E_k$ sera appelé élément amélioré).

Avantages et inconvénients de la méthode des centres mobiles La méthode des centres mobiles a l'avantage d'optimiser un critère simple de dispersion : l'inertie de la partition. Par ailleurs, la méthode converge toujours et ce, quelle que soit la partition initiale [Diday, 1971]. La pratique montre que cette convergence est très rapide, même dans le cas de grands jeux de données. On démontre que le critère d'inertie de la partition finale atteint un optimum *local* qui dépend en partie de la répartition initiale [Diday, 1971]. Lorsque les groupes ont une direction préférentielle (nuage de forme allongée), parallèle à celle des axes, un effet *ping-pong*[32] peut dégrader la classification. En général, il est préférable d'utiliser la méthode des nuées dynamiques plutôt que la méthode de Forgy, car cette dernière délimite moins bien les classes (les frontières sont plus franches et plus rectilignes) et elle est plus sensible à l'effet *ping-pong*. Cependant, pour certaines configurations du nuage de points, la méthode de Forgy permettra une meilleure partition que la méthode des nuées dynamiques (voir les quatre classes de *M. tuberculosis* H37Rv p. 218). Quand une méthode de partitionnement ne fonctionne pas sur un nuage de points particulier, on distingue clairement que la séparation des classes est artificielle : les frontières entre les classes sont rectilignes et verticales, lorsque les points sont représentés sur les deux premiers axes de l'AFC et colorisés en fonction de leur classe.

La méthode des centres mobiles possède deux inconvénients :

1. le nombre de groupes doit être connu *a priori*
2. la partition obtenue dépend du choix des centres initiaux

Pour s'affranchir de ces limites, une première solution consiste à procéder à une itération de la méthode des centres mobiles. En effet, la rapidité des calculs permet de comparer les différentes façons de répartir les objets en k groupes. On constate alors que plusieurs partitions possibles, et tout à fait acceptables, peuvent être de qualité voisine (Fig. 3.8 p. 144), ce qui représente à la fois un avantage et un inconvénient pour l'utilisateur. Effectivement, il n'existe pas une unique partition pour un jeu de données (en particulier, à quel groupe doivent être rattachés les individus frontaliers ?). Pour

[32] Les points sont à égale distance de deux centres.

palier cet inconvénient du choix des centres initiaux, on applique m fois l'algorithme des centres mobiles avec la configuration initiale où les k centres sont tirés au hasard. Parmi les m partitions obtenues, la meilleure est conservée d'après un critère de qualité. Il est recommandé d'effectuer au moins 20 essais (20 tirages aléatoires des centres initiaux) pour avoir une vue d'ensemble des classifications possibles [Delorme & Henaut, 1988]. Il est bien sûr possible d'itérer n'importe quelle variante de la méthode des centres mobiles : K-means itéré ou algorithme de *Forgy itéré*.

Ether est un programme de partionnement automatique par la méthode des centres mobiles itérée [Volle, 1997], intégré dans le module *GenoBool* de la plate-forme *Genostar* (voir p. 156). L'utilisateur doit choisir :

1. les données à classer (les coordonnées des points sur les trois premiers axes de l'AFC)
2. la variante de Forgy (*centroïds*) ou la variante des nuées dynamiques (*kernels*)
3. le nombre maximum de classes ($class\ number = 3$)
4. le nombre d'essais ($trials = 100$)
5. le seuil de la différence de qualité entre deux itérations pour le critère d'arrêt (*heterogeneity threshold*, $\sum_k R^{(r+1)} - \sum_k R^{(r)} < 0,001$)

Pour la méthode des nuées dynamiques, l'utilisateur doit en plus choisir la taille des noyaux (par exemple $kernel\ size = 25$). La partition finale conservée parmi tous les essais est celle qui minimise $\sum_k R^{(arret)}$.

En 1971, Diday avait prévu non seulement d'itérer son algorithme des nuées dynamiques, mais aussi de résoudre le problème du nombre de groupes connu *a priori* : la méthode des nuées dynamiques itérée permet de confronter m partitions et de rechercher des formes intéressantes s'agrégeant en leur cœur et permettant ainsi de les reconnaître. Ainsi pour résoudre le problème du nombre de groupes connu *a priori*, nous allons présenter d'autres techniques de classification.

Formes fortes La meilleure façon de tirer parti de la pluralité des solutions acceptables obtenues par la méthode des nuées dynamiques itérée est l'examen des formes fortes. Il s'agit de sélectionner les objets qui sont toujours classés ensemble dans les m partitions obtenues par les centres mobiles. Les formes fortes sont constituées des sous-ensembles d'objets qui ont toujours été réunis dans le même groupe final au cours des différents essais de partitions initiales [Delorme & Henaut, 1988]. Elles représentent donc des groupes homogènes et mettent en relief les objets d'attribution indécise qui n'appartiennent à aucune forme forte. L'étude du nombre de formes fortes permet d'éliminer le problème que pose le choix *a priori* du nombre de classes, puisque le nombre de formes fortes ne dépend pas directement du nombre de groupes k.

Puisque nous avons plusieurs partitions équivalentes et aucun critère objectif pour en choisir une plutôt qu'une autre, il est possible de s'intéresser au nombre de fois où deux objets ont été classés ensemble au cours de m partitions. Pour cela, on construit une matrice de co-occurrence. C'est une matrice carrée symétrique (de dimension $n * n$) qui comptabilise le nombre de fois x_{ij} où deux individus i et j sont classés ensemble parmi les m partitions. Par exemple, si $m = 20$ et

$x_{ij} = 20$, alors les objets i et j ont été classés vingt fois sur vingt ensemble. x_{ij} est l'inverse d'une distance ultramétrique[33]. A partir de cette matrice, on établit une représentation symbolique sous forme d'arbre[34] ou de cimetière américain (symboles + et -) qui permet de déduire le nombre de formes fortes (FIG. 3.8 C p. 144). La représentation symbolique ainsi construite possède les qualités suivantes :

- Elle est robuste, c'est-à-dire qu'elle n'est pas sensible à de petites variations du tableau de distances de départ.
- Elle est exempte de l'effet de chaîne, ce qui la rend pratiquement insensible à la suppression d'un objet. Autrement dit, les formes fortes résistent bien aux aléas statistiques : dans la figure 3.8 p. 144, la répartition en quatre formes fortes et deux objets inclassables reste inchangée quand on supprime l'objet d.
- Les objets qui s'agglomèrent de façon *hiérarchique*, définissant des groupes de plus en plus vastes, et ceux dont les affinités sont incertaines (à mi-chemin entre deux groupes par exemple) sont clairement mis en évidence. L'utilisateur voit ainsi, de manière objective, quels sont les objets sur lesquels il doit porter son attention au cours de l'interprétation des groupes.

L'étude détaillée de l'évolution des formes fortes en faisant varier k (de 2 à 10), permet de déterminer la façon dont les classes s'imbriquent les unes dans les autres. Cette exploitation de la méthode des formes fortes s'applique plus dans un cadre de phylogénie (évolution de la classification d'un individu dans une population selon le nombre de classes k) que dans un cadre de l'étude de l'usage des codons synonymes (détermination du nombre de formes fortes). Dans ce cadre, un niveau d'itération supérieur pour k est ici simplement ajouté à la méthode des nuées dynamiques itérée [Delorme & Henaut, 1988]. On obtient une matrice de co-occurrences et une représentation symbolique pour chaque k. On peut donc vérifier la congruence du nombre de formes fortes en fonction des différents k. Dans l'exemple de la figure 3.8 p. 144, on obtient quatre formes fortes et deux objets inclassables pour tout k compris entre 2 et 10.

Application à l'étude de l'usage du code génétique Selon les auteurs, le partitionnement des gènes en fonction de l'usage du code génétique est réalisé directement avec les données *brutes* (fréquences absolues ou relatives, RSCU, etc. ; [Mathe *et al.*, 1999, Hayes & Borodovsky, 1998b]) ou avec les coordonnées des individus sur les axes factoriels (ACP, AFC, etc. ; [Médigue *et al.*, 1991, Moszer *et al.*, 1999]). Dans la première approche, une méthode multifactorielle peut être utilisée, soit avant le partitionnement pour choisir le nombre de classes k, soit après le partitionnement pour obtenir une représentation graphique des résultats (nuage de points). Le partitionnement effectué sur les données brutes permet de prendre en compte un maximum d'information. La seconde approche présente plusieurs avantages. Tout d'abord, les méthodes multifactorielles réduisent le nombre de variables et les calculs de partitionnement sont donc plus simples et plus rapides. De

[33] Une distance ultramétrique est un score sans unité entre deux objets qui respecte l'inégalité triangulaire. Dans notre exemple, c'est un score de similitude (inverse d'une distance).

[34] La représentation en arbre n'est pas issue d'une classification hiérarchique sinon tous les objets seraient reliés par des branches dichotomiques.

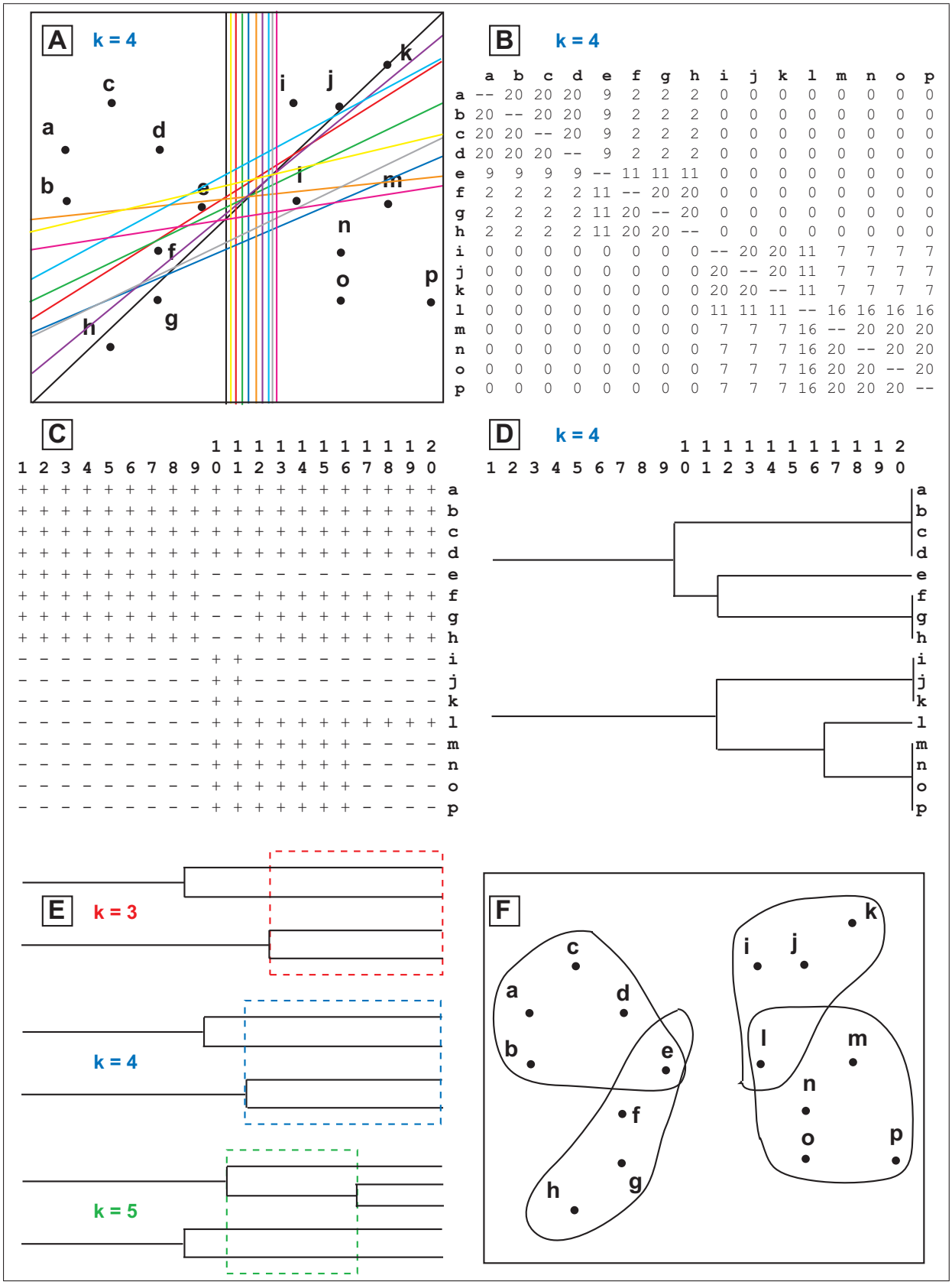


FIG. 3.8 – Principe de la recherche de formes fortes dans des multipartitions

FIG. 3.8 - Principe de la recherche de formes fortes dans des multipartitions

A) Exemple de Sneath et Sokal [Sneath & Sokal, 1973] sur lequel on réalise 10 fois l'algorithme des centres mobiles (10 tirages aléatoires des 4 centres initiaux). On obtient 10 partitions en 4 groupes. L'intersection de deux droites définit une partition en 4 groupes (il y a une couleur par partition). En pratique, 20 essais sont effectués (et non pas 10).

B) La matrice de co-occurrences indique le nombre de fois où deux objets sont classés ensemble sur 20 essais. Il y a 4 formes fortes (abcd, fgh, ijk et mnop) et deux objets inclassables (e et l). L'objet e est dit inclassable car il a été classé neuf fois sur vingt avec le groupe abcd et onze fois sur vingt avec le groupe fgh. Il y a différentes représentations équivalentes de cette matrice de formes fortes.

C) Représentation symbolique de la matrice de co-occurrences (cimetière américain). Cette représentation doit être lue colonne par colonne et de la droite vers la gauche. Un groupe est représenté par une série continue de signes + ou -. Les individus qui forment un groupe ont été regroupés *au moins* i fois sur vingt avec *au moins* un des individus de ce groupe (i correspond au numéro de colonne). Pour trouver le nombre de formes fortes, on compte le nombre de groupes (nombre de paquets de signes soit + soit -) pour chacune des i fois où deux objets ont été classés ensemble. Pour déterminer le nombre de formes fortes, on retient la meilleure plage qui réponde à deux critères : sa stabilité (sa longueur) et sa qualité (plus les i de cette plage sont élévées, meilleure est sa qualité). Dans cet exemple, les troisième et quatrième plages sont à la fois stables et de bonne qualité (de 12 à 16 dans la troisième et de 17 à 20 dans la quatrième), on en déduit quatre formes fortes : $\{abcd\}\{fgh\}\{ijk\}\{mnop\}$ et e, l sont des objets difficilement classables.

D) Représentation arborescente de la matrice de co-occurrences.

E) Représentation arborescente des matrices de co-occurrences pour une multi-partition où k=3, 4 et 5. Sur ces arbres les branches représentant les individus inclassables ont été supprimées. On obtient quatre formes fortes pour ces trois multipartitions.

F) Représentation des quatre formes fortes sur l'exemple de Sneath et Sokal (les classes sont recouvrantes).

plus, elles éliminent le bruit (variabilité résiduelle non structurée des données) et les partitions seront donc plus stables. Enfin, cette approche permet de confronter les résultats d'ordination des individus obtenus par les méthodes multifactorielles, aux résultats de regroupement obtenus par les méthodes de partitionnement.

Classification mixte ou combinée

Afin de pallier les inconvénients des méthodes de partitionnement (optimum local, k) et de classification hiérarchique ascendante (nombre d'individus), une solution consiste à combiner ces méthodes. La méthode de combinaison classique [Baccini & Besse, 2002, Lebart *et al.*, 2000] se réalise de la manière suivante :

1. Exécuter une méthode de partitionnement par réallocation dynamique en demandant un grand nombre de classes, de l'ordre de 10% du nombre total de points à classer. Par exemple, si le nombre de gènes est égal à 4000, on peut demander $k = 400$ groupes. Cela génère une partition, en partie arbitraire (les 400 groupes n'ont pas de sens biologique), mais qui permet de résumer le nombre d'individus.
2. Sur les centres de gravité des classes précédentes, une classification hiérarchique est réalisée afin de déterminer un nombre optimal de k classes.
3. Une méthode de partitionnement est alors exécutée sur tout l'ensemble, en fixant k le nombre de classes, et en choisissant les noyaux initiaux au centre des classes de l'étape précédente.

La méthode de la partition centrale est un autre exemple de classification mixte (logiciel SICLA ; [Croquette & Carlier, 1999]). Elle vise à obtenir une partition consensus, qui soit assez proche de l'ensemble des partitions de la multipartition de la façon suivante :

1. On définit une multipartition (un ensemble de m partitions effectuées sur le même ensemble de n individus avec k fixé). On recherche les formes fortes avec une définition stricte : une forme forte d'une multipartition est un ensemble d'individus *toujours* classés ensemble dans les m partitions.
2. On élimine les formes fortes d'effectif trop faible (par exemple $n \leq 2$), puis un tableau de distances entre formes fortes est calculé. Cette distance correspond au nombre de partitions pour lesquelles les deux formes fortes ne sont pas dans une même classe.
3. Une classification hiérarchique ascendante (méthode du saut minimum ; annexe E p. 409) est réalisée à partir du tableau de distances des formes fortes restantes.
4. Dans l'arbre de partitions ainsi obtenu, on recherche celle qui joue le rôle de partition centrale. On appelle partition centrale, la partition de la hiérarchie de formes fortes, la plus proche, en moyenne, des partitions de la multipartition (au sens de la distance entre partitions [Croquette & Carlier, 1999]).
5. Les formes fortes éliminées sont ensuite regroupées avec des classes existantes, ou encore regroupées en une classe *rebut*.

La classification hiérarchique ascendante, la méthode des formes fortes et la classification mixte ont l'avantage de ne pas fixer le nombre k de classes *a priori* ; cependant elles s'appuient sur des arbres (suite de partitions emboîtées) qui soulèvent le problème du choix de la partition optimale. En d'autres termes, à quel endroit doit on couper l'arbre ? [Lebart *et al.*, 2000]

Ainsi nous venons de voir la complémentarité entre analyse factorielle et classification. L'AFC peut être utilisée comme une étape préalable à la classification pour ses pouvoirs de description et de filtrage. L'utilisation conjointe de ces deux familles de méthodes peut mettre en évidence des groupes d'individus réunis par des facteurs latents inattendus. Les deux techniques se valident mutuellement [Lebart *et al.*, 2000]. Les classes de gènes mises en évidence peuvent alors être prises en compte lors de la prédiction de gènes, chaque groupe étant alors utilisé pour construire un modèle de gènes homogènes dans leur usage des codons synonymes (voir p. 197).

Chapitre 4

Annotation de génomes : quels moyens informatiques ?

Pour annoter un génome procaryote, nous disposons de données organisées dans des banques et des bases, et de méthodes d'analyse implémentées dans des programmes informatiques. Ce chapitre est destiné à introduire les concepts d'intégration de données, de méthodes, puis de bases de connaissances, concepts qui sont la clé de voûte des plates-formes bioinformatiques discutées dans la seconde section de ce chapitre.

4.1 Intégration d'outils distribués et hétérogènes

Les outils présentés dans les chapitres précédents répondent, le plus souvent, à des objectifs différents. Il est illusoire de chercher à concevoir un outil universel permettant de répondre à plusieurs questions biologiques. Il faut accepter la pluralité des problématiques, qui implique nécessairement une pluralité des outils. Aussi, le problème informatique majeur de l'annotation des génomes procaryotes est lié à l'existence d'une multitude d'outils distribués fournissant des données hétérogènes, tant en termes de nature que de format. C'est pourquoi depuis le milieu des années 90, le mot clé en bioinformatique est devenu « intégration ».

4.1.1 Standards facilitant les compatibilités entre données de banques et de bases

Nous avons vu précédemment que le format de données des banques actuelles n'est pas toujours identique (Swiss-Prot, GenPept, PFam ...). L'homogénéisation des formats relève d'un problème essentiellement technique (extraction des données des fichiers des banques), et permet d'assurer la compatibilité syntaxique des données. Plus complexe est le problème d'homogénéisation des concepts représentant les entités et des valeurs prises par les attributs. Par exemple, nous avons vu que le qualificatif *gene* des banques de séquences nucléiques, désigne un concept ambigu pour lequel il existe différentes interprétations (voir p. 65). Bien que des standards aient été développés

pour faciliter la gestion de ces problèmes, seuls des experts sont capables de trouver des solutions. Aujourd'hui, trois standards de fonctionnalités complémentaires facilitent la mise en œuvre de solutions définies par des experts afin d'assurer la compatibilité syntaxique des données :

- CORBA[1] (*Common Object Request Broker Achitecture*)
- XML[2] (*eXtended Markup Langage*)
- UML[3] (*Unified Modeling Langage*)

Name	ST	Description	Availability	Reference
DiscoveryLink	F	Middleware system based on a virtual relational database.	IBM	Haas-2001
Ensembl	W	Software system integrating eukaryotic genomic data and bioinformatics tools.	The Wellcome Trust Sanger Institute/European Bioinformatics Institute (EBI)	Hubbard-2002
Entrez	W	Information retrieval system based on a relational database.	NCBI	Schuler-1996
GenoMax	W	Enterprise-level integration of bioinformatics tools and data sources. Data are stored within a relational database.	InforMax	http://www.informaxinc.com/solutions/genomax/index.html
Kleisli	F	Mediator system encompassing a nested relational data model, a high-level query language and a powerful query optimizer.	geneticXchange	Chung-1999
SRS	F	Indexed flat-file system built on the model of a document retrieval system.	EBI/Lion biosciences	Etzold-1996
TAMBIS	F	Retrieval-based information system using an ontology for molecular biology and bioinformatics.	University of Manchester, UK	Stevens-2000
XML-based system	N/A	Tagged text files system based on a XML Schema.	W3C	http://www.w3.org/XML/

ST system type; **F** federated database; **W** data warehouse; **XML** extensible markup language; **N/A** not applicable.

TAB. 4.1 – Sytèmes d'intégration de donnéesCe tableau est extrait de la publication [Durand *et al.*, 2003] .

Deux approches permettent d'intégrer des données de différentes sources, dans un contexte favorable à leur compatibilité syntaxique et sémantique, afin de les rendre accessibles sous forme d'une base unique.

L'approche fédérative consiste à ajouter, au-dessus des bases existantes, une couche logicielle qui offre les interfaces nécessaires entre les bases faisant ainsi apparaître l'ensemble comme une seule base virtuelle (par exemple SRS [Etzold *et al.*, 1996]). L'approche entrepôt de données consiste à agréger, au sein d'un schéma unique (UML), les données (XML) de différentes bases (*datawarehouse* ; TAB. 4.1 p. 150). Ces données sont organisées par sujets et gérées dans un environnement de stockage spécialisé. L'approche entrepôt possède plusieurs avantages par rapport à l'approche fédérative. Les performances de temps d'accès aux données et de traitement des requêtes sont théoriquement meilleures ; en contre partie, cette approche nécessite des mises à jour très régulières. L'ontologie est implicite dans le schéma conceptuel commun à toutes les données réunies. Ainsi, l'approche entrepôt offre un contexte plus favorable à la compatibilité sémantique.

[1]Les spécifications de CORBA 2.0 `http://www.omg.org` ont été adoptées en 1994.
[2]XML`http://www.xml.com/` a été créé en 1999.
[3]Fin 1997, UML`http://www.omg.org/uml/` qui est devenu une norme OMG (*Object Management Group*).

4.1.2 Bases de connaissances et intégration de méthodes

Le concept de connaissances est lié à l'interprétation de données compatibles. Cette interprétation repose sur le mécanisme d'inférence (génération de nouveaux faits à partir des données).

Afin de garantir l'interopérabilité syntaxique et sémantique des données à l'intérieur d'une base (ou entre deux bases), il faut passer de la notion de base (ou entrepôt) de données, à celle de base (ou entrepôt) de connaissances. Les problèmes de compatibilité syntaxique et sémantique des données doivent être résolus au niveau des schémas conceptuels [Morgat & Rechenmann, 2002]. La représentation de connaissances permet l'interopérabilité de plusieurs bases de données de domaines biologiques qui se recouvrent, et la modélisation explicite de réseaux biologiques par le biais des associations et des classes.

Pour gérer l'inférence de connaissances, le modèle doit être extensible afin d'accueillir de nouvelles connaissances (données inférées qui peuvent révéler la nécessité d'importer de nouvelles données primaires, etc.). Aussi, le second concept clé des bases de connaissances est celui de l'inférence. Les deux mécanismes d'inférence majoritairement utilisés sont l'induction et la déduction. Classiquement, en bioinformatique, l'inférence de connaissance par induction est basée sur l'utilisation d'un ensemble de méthodes, ce qui nécessite d'intégrer des programmes d'analyses dans un système assurant la compatibilité des données en entrée et en sortie.

Le moyen le plus simple d'*inférer des connaissances* est d'avoir, à sa disposition, à la fois un ensemble de données structurées et un ensemble de méthodes intégrées. A l'opposé du problème de l'intégration des données dans un environnement homogène, celui de l'intégration des méthodes a suscité, à ce jour, un intérêt moindre. En effet, les biologistes qui ne travaillent pas à *grande échelle* (sur de grands ensembles de données à la fois), se satisfont des méthodes distribuées en ligne. Les bioinformaticiens, quant à eux, vont rapatrier et compiler les méthodes dont ils ont besoin (ou bien même développer celles qui ne sont pas disponibles) pour les enchaîner au sein d'un script. Le manque de synchronisation entre les différents développements réalisés par la communauté de bioinformaticiens génère une redondance des méthodes et une perte de temps indéniable. Il semble donc crucial de pouvoir disposer d'un environnement permettant au programmeur d'accéder à toutes les méthodes disponibles au sein d'un groupe de travail, environnement qui faciliterait aussi la compréhension du rôle de ces méthodes.

Le *package*, ou « boîte à outils », est un logiciel qui intègre différentes méthodes. Il est parfois possible de construire des stratégies d'analyse qui permettent d'enchaîner les méthodes à partir d'un langage *ad hoc*. On peut citer les packages *Genetics Computer Group* (*GCG*), *European Molecular Biology Open Software Suite* (*EMBOSS*), etc. (Tab. 4.1 p. 154). Les packages peuvent être couplés à des SGBD, mais sans intégration réelle des deux systèmes dans un seul et même environnement. Bien que le package soit une avancée dans la méthodologie informatique, il est aujourd'hui évident que les données d'une part, et les méthodes d'autre part, sont indissociables : les données biologiques sont utilisées en entrée de programmes qui génèrent eux même de nouvelles données à stocker. Enfin, l'efficacité d'une méthode repose généralement sur un ensemble de paramètres dont les valeurs sont

souvent estimées sur des jeux d'apprentissage (des données biologique). C'est pourquoi une des plates-formes présentées dans la section qui suit (*Genostar*) repose sur une structure de base qui intègre à la fois les données et les méthodes d'analyse.

4.2 Plates-formes d'annotation et d'exploration des génomes procaryotes

Au milieu des années 90, devant l'augmentation exponentielle du nombre de séquences, il devenait tentant de développer et d'utiliser des méthodes d'analyse automatiques. Ainsi une première catégorie de plates-formes d'annotation met en œuvre des pipelines d'analyses automatiques, l'interaction avec l'utilisateur étant alors minimale. Les résultats de l'annotation fonctionnelle des gènes prédits sont présentés dans des pages HTML. Dans cette catégorie, on peut citer le logiciel *GeneQuiz* ou encore *Magpie* [Andrade *et al.*, 1999, Gaasterland & Sensen, 1996]. L'annotation strictement automatique des séquences génomiques n'étant ni facile, ni fiable, de nombreux environnements plus interactifs sont aujourd'hui très utilisés : c'est le cas des plates-formes *semi-automatiques Manatee* (Tab. 4.1 p. 154), *ERGO* [Overbeek *et al.*, 2003], *GeneDB* [Hertz-Fowler *et al.*, 2004] qui offrent, en complément d'un pipeline d'analyses, des interfaces d'annotations manuelles permettant de valider les annotations automatiques. Enfin, *Artemis* est avant tout une excellente interface graphique dédiée à l'annotation strictement manuelle des séquences génomiques [Rutherford *et al.*, 2000]. Comparés aux environnements automatiques, ces systèmes offrent généralement des modèles de données biologiques plus sophistiqués (ils reposent sur des SGBDR ou SGBDO), et/ou des interfaces graphiques interactives permettant d'extraire, d'analyser, d'annoter, de modifier, de visualiser les données représentées par ces modèles. A titre d'exemple, nous décrirons brièvement les plates-formes *GeneQuiz* (automatique), *Manatee* (semi-automatique), et *Artemis* (manuelle). Nous terminerons cette section en présentant une plate-forme plus récente dédiée à l'annotation et à l'exploration de séquences génomiques : *Genostar*.

4.2.1 *GeneQuiz*

GeneQuiz [Andrade *et al.*, 1999] est une plate-forme d'annotation fonctionnelle automatique accessible en ligne qui n'utilise que des méthodes d'analyse de séquences protéiques. L'utilisateur fournit une séquence protéique, et récupère une liste d'annotations fonctionnelles.

GeneQuiz est composé de quatre modules : GQUpdate, GQSearch, GQreason et GQbrowse. GQUpdate gère la mise à jour quotidienne de banques de séquences, de motifs et de structures (SWALL, PROSITE, PDB). Sur chaque séquence protéique requête le module GQSearch est chargé d'exécuter les méthodes de comparaison du type *Blastp*, *Fasta*, recherche de motifs, mais aussi recherche d'homologie dans la banque de structure 3D. Puis le module GQreason filtre les protéines sélectionnées par GQSearch en fonction de leur description. Ce module fait une analyse syntaxique du champ de description afin de vérifier si l'information contenue est valide ou non. Lorsque la

description est acceptable, les seuils prédéfinis sur le score et la E-value des programmes *Blast* et *Fasta* permettent de ranger la fonction de chaque protéine requête dans différentes catégories (*clear*, *tentative*, *marginal* et *unknown*). Finalement, le module GQbrowse est chargé de synthétiser les résultats obtenus pour chaque méthode dans une page HTML.

L'attribution fonctionnelle automatique développée par *GeneQuiz* combine de façon astucieuse plusieurs résultats d'analyse afin d'attribuer une fonction unique à la protéine analysée. *GeneQuiz* a servi notamment à l'analyse du génome de *M. jannaschii* pour la prédiction de fonctions protéiques [Andrade *et al.*, 1997]. Elles se heurtent cependant inévitablement au problème de l'accumulation des erreurs d'annotation dans les banques de séquences, mais aussi à l'organisation souvent modulaire des protéines, conduisant alors à des annotations soit « fausses », soit incomplètes.

4.2.2 *Manatee*

Manatee (*Manual Annotation Tool Etc, Etc*) est une plate-forme d'annotation semi-automatique de génomes procaryotes developpée en langage Perl au TIGR[4] (Tab. 4.1 p. 154). Le moteur d'annotation automatique, décrit dans différentes publications [Nierman *et al.*, 2001, Tettelin *et al.*, 2001], est constitué d'un pipeline externe. Les CDS sont prédites par le programme *GLIMMER* [Delcher *et al.*, 1999a], puis comparées aux séquences d'une base non redondante d'acides aminés constituée à partir de PIR-NREF et de la base *CMR* (*Comprehensive Microbial Resources*) du TIGR. Le programme *BER* (*Blast-Extend-Repraze*), utilisé à cet effet, permet aussi de repèrer des décalages du cadre de lecture. Les séquences protéiques prédites sont aussi examinées par le programme *HMMPfam* (prédiction de domaines protéiques). Le programme *AutoAnnote*, qui analyse les résultats des recherches BER et HMMpfam, permet d'assigner à chaque protéine un nom commun, un symbole de gène, un numéro de la commission enzyme, et une classe fonctionnelle de *Gene Ontology* [Ashburner *et al.*, 2000]. Dans le cas des génomes procaryotes, le TIGR utilise la classification fonctionnelle de Monica Riley [Serres *et al.*, 2004]. Les résultats de l'annotation automatique sont stockés dans une base MySQL : ils sont visualisés et édités pour une annotation manuelle via le navigateur de *Manatee*. Ainsi, l'interface cartographique de *Manatee* permet d'identifier rapidement les gènes et d'assigner une fonction à partir des résultats de similitudes, de familles de paralogues et des suggestions d'annotation générées par les analyses automatiques. Le centre de séquençage du TIGR séquence et annote de nombreux génomes procaryotes comme *N. meningitidis* MC58 (serogroup B), *Caulobacter crescentus*, *Streptococcus pneumoniae* TIGR4, *Bacillus anthracis* [Tettelin *et al.*, 2000, Nierman *et al.*, 2001, Tettelin *et al.*, 2001, Read *et al.*, 2003].

4.2.3 *Artemis*

Parmi l'ensemble des génomes de micro-organismes aujourd'hui disponibles, la qualité des annotations est généralement supérieure chez les annotateurs qui utilisent des interfaces graphiques destinées à une validation des annotations fonctionnelles [Bocs *et al.*, 2002]. Le logiciel *Artemis*

[4] Accès libre aux sources, `http://www.tigr.org/tdb/mdb/mdbcomplete.html`

Name	Description	URL	Organisms	Nature of data	Data model (implementation)[1]	Architecture	Local installation[2]	Reference
GCG Wisconsin Package	Sequence analysis package.	*http://www.accelrys.com/products/gcg_wisconsin_package/index.html*	Prokaryotes Eukaryotes	DNA/RNA/protein sequences	Yes (RDBMS)	Client-server	Yes	-
EMBOSS	Open source software package for sequence analysis.	*http://www.hgmp.mrc.ac.uk/Software/EMBOSS*	Prokaryotes Eukaryotes	DNA/RNA/protein sequences	No	-	Yes	Rice-2000
Darwin	Open source software package for sequence comparisons and phylogenetic tree building.	*http://cbrg.inf.ethz.ch/Darwin*	Prokaryotes Eukaryotes	DNA/RNA/protein sequences	No	-	Yes	Gonnet-2000
PEDANT	System for completely automated and exhaustive analysis of protein sequence sets.	*http://pedant.gsf.de/*	Prokaryotes Eukaryotes	DNA/protein sequences Functional predictions Protein-protein interactions	Yes (RDBMS)	Client-server	Web server	Frishman -2001
GeneQuiz	Fully automated system for genome analysis.	*http://jura.ebi.ac.uk:8765/*	Prokaryotes Eukaryotes	Protein sequences Functional predictions	Yes (RDBMS)	Client-server	Web server	Andrade -1999
MAGPIE/ EGRET	Fully automated sequence for genome analysis.	*http://genomes.rockefeller.edu/ magpie/*	Prokaryotes (EGRET for Eukaryotes)	DNA/protein sequences	No	Client-server	Yes	Gaasterland-1996 Gaasterland-2000
Biofacet (LASSAP)	Software platform for comparative genomics.	*http://www.gene-it.com*	Prokaryotes Eukaryotes	DNA/protein sequences	No	-	Yes	Glemet-1997
SEALS	Software package for large-scale, semi-automated sequence analysis.	*http://www.ncbi.nlm.nih.gov/CBB research/Walker/SEALS/index. html*	Prokaryotes Eukaryotes	DNA/protein sequences	No	-	Yes	Walker-1997
ASAP	A Systematic Annotation Package for community analysis of genomes.	*https://asap.ahabs.wisc.edu/ annotation/php/ASAP1.htm*	Prokaryotes	DNA/protein sequences Experimental data	Yes (RDBMS)	Client-server	Yes	Glasner-2003
AceDB	Database system with many specific displays and tools for genomic data.	*http://www.acedb.org*	Prokaryotes Eukaryotes	DNA/RNA/protein sequences Physical and genetic map	Yes (OODB)	Client-server	Yes	-
Artemis	Genome viewer and annotation tool.	*http://www.sanger.ac.uk/Software/Artemis/*	Prokaryotes Eukaryotes	DNA sequences	No	-	Yes	Rutherford-2000
Genotator	Genome viewer and sequence annotation tool.	*http://www.fruitfly.org/~nomi/genotator*	Eukaryotes	DNA sequences	No	-	Yes	Harris-1997
Sequin	Sequence viewer and annotation tool, primarily designed for sequence submission.	*http://www.ncbi.nlm.nih.gov/Sequin*	Prokaryotes Eukaryotes	DNA/protein sequences	Yes (NCBI data model/ASN.1)	Client-server	Yes	-
Apollo	Genome annotation editor.	*http://www.fruitfly.org/annot/apollo/*	Eukaryotes	DNA sequences	Yes (RDBMS; Gadfly/Ensembl projects)	Client-server	Yes	Lewis-2002
Manatee	Web-based gene evaluation and genome annotation tool.	*http://manatee.sourceforge.net/*	Prokaryotes Eukaryotes	DNA/protein sequence Functional prediction outputs	Yes (RDBMS)	Client-server	Yes	-
ERGO	Software system for a comprehensive analysis of genes and genomes.	*http://ergo.integratedgenomics.com/ERGO*	Prokaryotes Eukaryotes	Genomic data Expression data Regulatory data	Yes (RDBMS)	Client-server	Web server	Overbeek-2003
UberTool	Software system for the integration and analysis of molecular biological data.	*http://www.science-factory. com/products.html*	Prokaryotes Eukaryotes	DNA/protein sequences 3D structure Expression data	Yes (OODB)	Client-server	Yes	-
Genostar	Software platform for genome annotation and exploration.	*http://www.genostar.org*	Prokaryotes	DNA sequences Experimental data	Yes (OODB)	-	Yes	-

[1]Does the system rely on an explicit data model, and if yes, what is the implementation? **RDBMS** relational database management system; **OODB** object-oriented database. A hardwired model (eg, Java classes) is not considered as explicit. [2]Does the system require a local installation, or is it only accessible through a web server?

FIG. 4.1 – A) Plate-forme d'analyse de génomes (côté données)

Name	Nature of methods	Built-in annotation pipeline	User control execution[1]	User control strategy[2]	Data visualization[3]	Data edition[4]
GCG Wisconsin Package	Sequence analysis tools.	No	GUI or command line	Yes (programming)	Passive	Manual
EMBOSS	Complete set of sequence analysis tools.	No	GUI or command line	Yes (programming)	Active	Manual
DARWIN	Sequence comparison algorithms and phylogenic trees reconstitution.	No	Command line	Yes (programming)	Passive	Manual
PEDANT	Tools for protein function and structure prediction, gene context exploration (SNAP method).	Yes	Command line	No	Active	GUI
GeneQuiz	Complete set of protein annotation tools for automatic functional assignment.	Yes	GUI HTML	No	Active	No
MAGPIE/EGRET	Sequence annotation tools for automatic functional assignment.	Yes	Command line	Yes (programming)	Active	GUI
Biofacet (LASSAP)	Sequence comparison algorithms (eg, local, global or blast).	Yes	Command line	Yes (programming)	Active	Manual
SEALS	Sequence analysis tools.	No	Command line	Yes (programming)	Passive	Manual
ASAP	Sequence annotation tools.	No	No (external analysis tools)	No	Active	Yes (annotator and curator levels only)
AceDB	Genome annotation and visualization tools.	No	GUI or command line	No	Active	GUI
Artemis	Genome annotation and visualization tools.	No	GUI	No	Active	GUI
Genotator	Genome annotation and visualization tools.	Yes	GUI or command line	No?	Active	GUI
Sequin	Sequence annotation tools.	No	GUI	No	Active	GUI
Apollo	Genome annotation and visualization tools.	No	No	No	Active	GUI
Manatee	Genome annotation and visualization tools.	Yes	No	No	Active	GUI
ERGO	Genome annotation and visualization tools. Methods for comparative genomics (eg, gene contexts, pathways or plylogenetic clusters).	Yes	No(the internal pipeline belongs to Integrated Genomics Inc)	No	Active	No
UberTool	Sequence, structure and expression data analysis.	Yes	GUI	Yes (GUI)	Active	GUI
Genostar	Genome annotation and visualization tools, gene context exploration.	Yes	GUI	Yes (programming)	Active	GUI

[1]Can the user control the execution of methods on data? [2]Can the end user design his/her own strategies? 'Programming' means that this is possible, by writing code (eg, Unix scripts or programming API). [3]Describes interactions with data system. Passive: display data only; active: dynamic (usually dedicated) interface connected to the internal representation of data. [4]How can the end user edit and annotate the data? 'Manual' means that edits are possible through direct edit of (text) files; 'GUI' (graphical user interface) means that dedicated graphical editors are available.

FIG. 4.1 – B) Plate-forme d'analyse de génomes (côté méthodes)

Ces tableaux sont extraits de la publication [Durand *et al.*, 2003]

développé au Sanger Centre (Tab. 4.1 p. 154) dispose d'une interface graphique très conviviale permettant, à partir des résultats de plusieurs méthodes ayant été exécutées au préalable, d'annoter chacun des objets génomiques caractérisés (CDS, introns et exons ...), et de les sauvegarder au format de la banque EMBL. Cette plate-forme ne repose pas sur l'utilisation d'un SGBD et les quelques méthodes accessibles au niveau de l'interface graphique restent rudimentaires (recherche d'ORFs, calcul de GC skew, ...). L'exécution des programmes *Blastp* et *Fasta* sur des CDS choisies est réalisée à l'extérieur *Artemis*. Ainsi, c'est avant tout l'interface graphique dynamique qui rend ce logiciel très attractif : il apporte en particulier beaucoup d'aide à l'annotation précise des codons d'initiation de la traduction.

Le centre de séquençage du Sanger[5] séquence et annote de nombreux génomes bactériens, en particulier des organismes pathogènes comme *M. tuberculosis* H37Rv, *N. meningitidis* Z2491 (Serogroup A), *Y. pestis* CO92 [Cole *et al.*, 1998, Parkhill *et al.*, 2000, Parkhill *et al.*, 2001b]. L'annotation est basée autant que possible sur celle de gènes et de protéines déjà caractérisés. Un programme dérivé d'*Artemis*, ATC, a récemment été développé dans le but de comparer les régions génomiques de plusieurs chromosomes bactériens [Parkhill *et al.*, 2003].

4.2.4 Genostar

Les exemples de plates-formes d'annotation que nous venons de voir séparent très nettement les méthodes d'analyse (pipeline), des données biologiques (séquences et résultats d'analyse), ces dernières étant le plus souvent organisées dans des bases de données. Quelle que soit son architecture, une plate-forme d'annotation et d'exploration de génomes devrait intégrer une base de données biologiques, une base de méthodes d'analyse de ces données et des interfaces utilisateurs qui permettent de gérer et d'interroger les données, de lancer et contrôler l'exécution de stratégies, et enfin de visualiser et d'annoter les résultats des méthodes. C'est dans cet esprit que le système coopératif d'aide à l'analyse de séquences procaryotes, *Imagene* [Médigue *et al.*, 1999a], puis plus récémment la plate-forme *Genostar* [Durand *et al.*, 2003], ont été développés.

La plate-forme de génomique exploratoire *Genostar* est composée de trois modules : *GenoAnnot* (annotation syntaxique), *GenoBool* (analyse multivariée) et *GenoLink* (annotation fonctionnelle). Un module repose sur deux services : *GenoCore*(représentation et gestion des connaissances liées au domaine traité par le module en question) et *GenoViews* (représentation graphique de ces connaissances). Un des points forts de *Genostar* réside dans la modélisation uniforme des connaissances liées à deux concepts : les faits (ou données biologiques) et les méthodes (ou programmes d'analyse). En effet, l'application *GenoCore* repose sur le système de représentation et de gestion de connaissances factuelles *AROM* [Page *et al.*, 2000] et sur le système de représentation et de gestion de connaissances méthodologiques *AROMTasks* (Fig. 4.2 p. 160).

Les classes de faits, les associations et les classes de problèmes sont organisées de manière hiérarchique (relation de spécialisation-généralisation). En pratique, la communication entre la base

[5] `http://www.sanger.ac.uk/Projects/Microbes/`

de connaissances factuelles et la base de connaissances méthodologiques est implicite puisque les entrées et les sorties des instances de méthodes sont des objets (instances de classe) et des tuples (instances d'association) de la base de connaissances. Par ailleurs, *Genostar* a l'avantage d'être modulaire et extensible. En effet, il est possible d'étendre l'ontologie du modèle en décrivant de nouvelles entités. De même, on peut facilement ajouter de nouvelles méthodes. Enfin, il est possible de développer de nouveaux modules dédiés à d'autres domaines biologiques spécifiques : par exemple, le module *GenoExpertBacteria* (*GEB*), développé au sein du groupe HELIX (INRIA), intègre les résultats des autres modules de la plate-forme, pour permettre une exploration des connaissances des gènes, des protéines et des voies métaboliques des génomes bactériens complets.

La plate-forme *Imagene* a été utilisée pour l'annotation des génomes de *B. subtilis* [Kunst *et al.*, 1997] et de *Mycoplasma pneumoniae* [Chambaud *et al.*, 2001] et *Genostar* est aujourd'hui utilisé pour l'annotation de *Ehrlichia ruminantium*[6]. La propriété d'interopérabilité des modules de *Genostar* est exploitée au sein de notre laboratoire dans le cadre d'une stratégie visant à affiner l'étape d'apprentissage pour la prédiction de gènes (voir p. 197).

4.3 Les projets HAMAP et HERBS

Nous ne pouvons pas terminer ce chapitre sans mentionner le travail de réannotation de génomes bactériens réalisé par le groupe de bioinformatique du SIB (Genève), en collaboration avec l'équipe HELIX de l'INRIA (Grenoble).

Le projet *HAMAP* (*High-quality Automated and Manual Annotation of microbial Proteomes* [Gattiker *et al.*, 2003]) a été mis en place pour faire face à l'avalanche de génomes bactériens complets produits par les centres de séquençage, en automatisant certaines tâches d'annotation des protéines nouvellement identifiées dans TrEMBL. Ce projet[7] recense actuellement 113 génomes procaryotes complets, codant au total plus de 300 000 protéines. Dans le cadre de ce projet, deux catégories de protéines essentiellement sont étudiées : les protéines qui ne présentent pas de similitudes avec les protéines de la SWALL (catégorie *Hypothetical protein*, *ORFan* ou *No similarity*) et les protéines qui appartiennent à des familles caractérisées, pour lesquelles des règles d'annotation peuvent être clairement définies (catégories *Belongs to a family* et *May belong to a family*). Les protéines qui n'appartiennent à aucune de ces deux catégories sont classées dans *Other similarity* et seront annotées ultérieurement par la méthode manuelle traditionnelle.

L'annotation d'une protéine de la catégorie *appartient à une famille* n'est pas simplement déduite de son appartenance à une famille, mais de toute une série de contrôles, tels que l'adéquation d'une séquence avec celles de la famille correspondante, ou encore la cohérence de la présence d'une protéine spécifique dans le contexte global du fonctionnement de l'organisme. Dans *HAMAP*, l'annotation automatique s'appuie sur une base de connaissances étendues des protéomes microbiens, qui inclut les caractéristiques fonctionnelles et structurelles, ainsi que les particularités génomiques

[6] `http://www.genostar.org/french/actualites/utilisation.htm`
[7] `http://ca.expasy.org/sprot/hamap/hamap_stat.html`

et les voies métaboliques d'un organisme. Ces connaissances peuvent être exploitées de manière systématique grâce à leur formalisation explicite et à leur mise à disposition sur un serveur web[8] dédié à l'étude des protéomes microbiens. Actuellement, environ 890 familles de protéines microbiennes orthologues, définies par des experts annotateurs sont regroupées dans la banque de familles *HAMAP*. Ainsi, si l'on s'intéresse par exemple à la glycolyse, toutes les familles des protéines intervenant dans ce processus vont être décrites. Cette banque de familles est utilisée pour annoter les nouvelles séquences protéiques de la catégorie *appartient à une famille*. Par exemple MF_00473 (*Microbial Family*) définit les règles et le profil caractérisant les séquences de la famille phosphoglucose isomérase.

Le pipeline d'annotation fonctionnelle est résumé dans la figure FIG. 4.3 p. 161. La première étape consiste à éliminer la redondance et à détecter les conflits (codons d'initiations alternatifs, décalages du cadre de lecture, identification de paralogues). La seconde étape répartit les protéines dans les différentes catégories en utilisant la collection de profils *HAMAP*. La propagation automatique des annotations, dans le cas d'une nouvelle séquence protéique de la catégorie *appartient à une famille* signifie qu'elle est en adéquation avec les règles caractérisant cette famille. Les problèmes rencontrés pour les séquences de la catégorie *peut appartenir à une famille* (conflits, longueur insuffisante) sont corrigés manuellement. La dernière étape traite le cas des protéines de la catégorie *pas de similitudes*. Elles sont analysées par différentes méthodes : recherche du peptide signal, de régions transmembranaires, de super hélices (*coiled-coil*), d'intéines (introns auto-catalytique de groupe II), de sites de liaison à l'ATP, etc. Le projet *HAMAP* génère ainsi des annotations automatiques et de haute qualité, qui couvrent 4 à 40% (*S. coelicolor* et *B. aphidicola* resp.) des protéomes.

Vérifier la cohérence globale de l'annotation d'un protéome consiste à examiner l'ensemble des protéines impliquées dans les différents processus biologiques qui régissent l'organisme microbien. Cette problématique nécessite de prendre en compte toutes les informations disponibles, aussi bien au niveau de l'organisme (*e.g.* physiologie, phylogénie) qu'au niveau des protéines (fonctions, intéractions). C'est dans cet esprit que le groupe Helix[9] apporte sa contribution au projet *HAMAP*. La base de connaissances factuelles est constituée de trois parties :

1. une base de données dédiée aux protéomes microbiens (faits à traiter)
2. une base de familles *HAMAP* (faits sûrs)
3. une base de règles définissant les processus biologiques, développée par le groupe Helix, en étroite collaboration avec les annotateurs de Swiss-Prot

Cette base de connaissances factuelles est associée à un moteur de règles nommé *HERBS*[10] (*HAMAP Expert Rules Based System*) développé à l'INRIA, et basé sur le système expert JESS[11] (*JAVA Expert Shell System*, langage de règles shell CLIPS). Le système expert *HERBS* a pour objectif de déceler des incohérences ou erreurs d'annotation. Ainsi, il est capable de détecter des

[8] `http://us.expasy.org/sprot/hamap/families.html`
[9] `http://www.inrialpes.fr/helix/SIB/sibelius.html`
[10] `http://www.inrialpes.fr/helix/SIB/herbs_janvier2003.pdf`
[11] `http://herzberg.ca.sandia.gov/jess/`

protéines absentes d'une voie métabolique donnée (*missing proteins*) et des protéines inattendues (*unexpected proteins*). Une interface graphique permet aux annotateurs de visualiser les protéines *missing* et *unexpected* de chaque voie de biosynthèse disponible d'un organisme donné. Actuellement testée par les annotateurs de Swiss-Prot, ce premier prototype permet déjà de valider ou d'invalider un certain nombre d'annotations. Ainsi, l'expertise des annotateurs est mise à profit pour développer des systèmes de règles permettant d'automatiser certaines étapes de l'annotation fonctionnelle et reste indispensable pour les autres étapes. Nous verrons qu'une partie des travaux de cette thèse s'intègre parfaitement au projet *HAMAP* dans le sens où ils apportent une évaluation de la qualité de prédiction des CDS qui varie selon les centres de séquençage (voir p. 327).

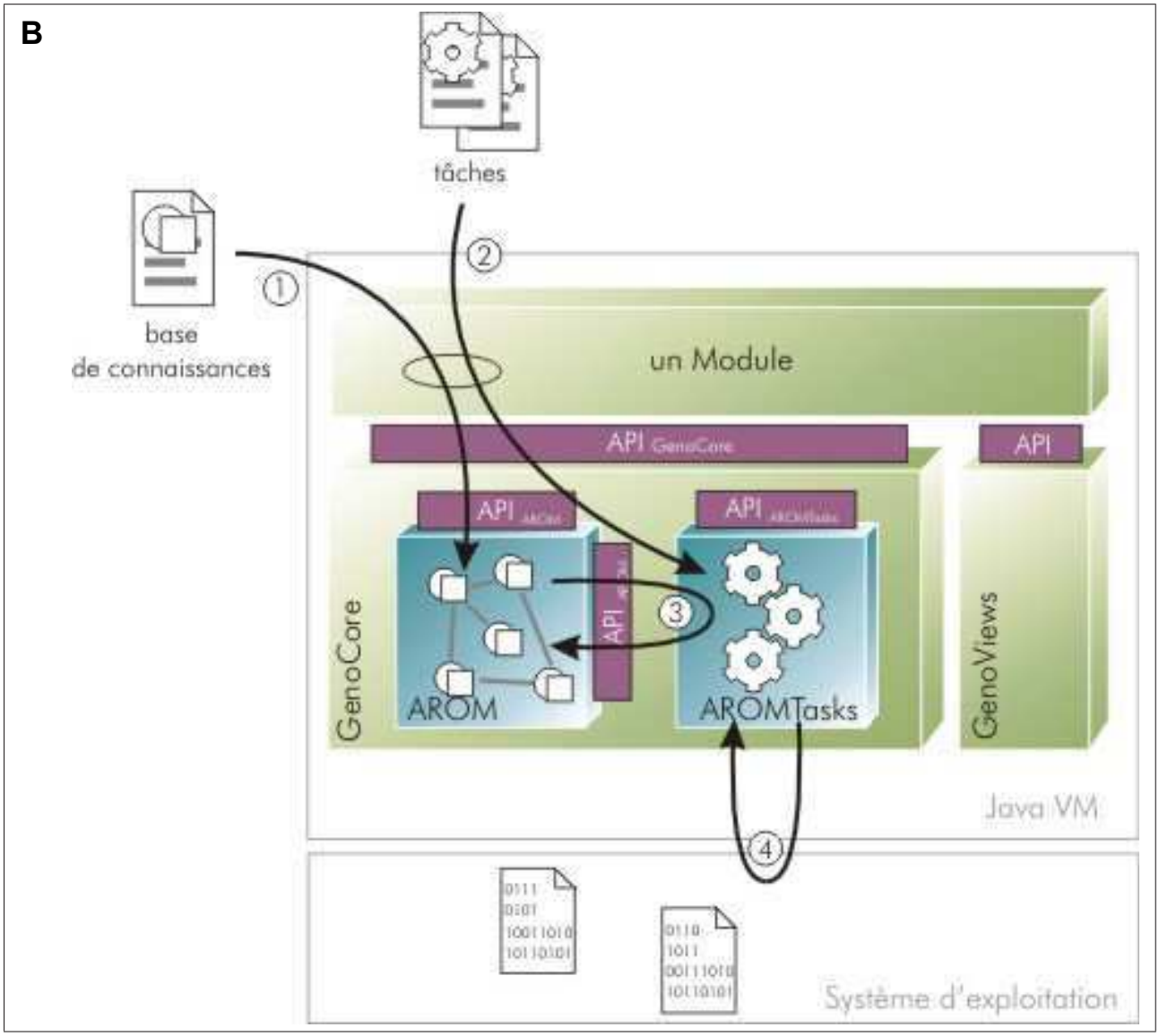


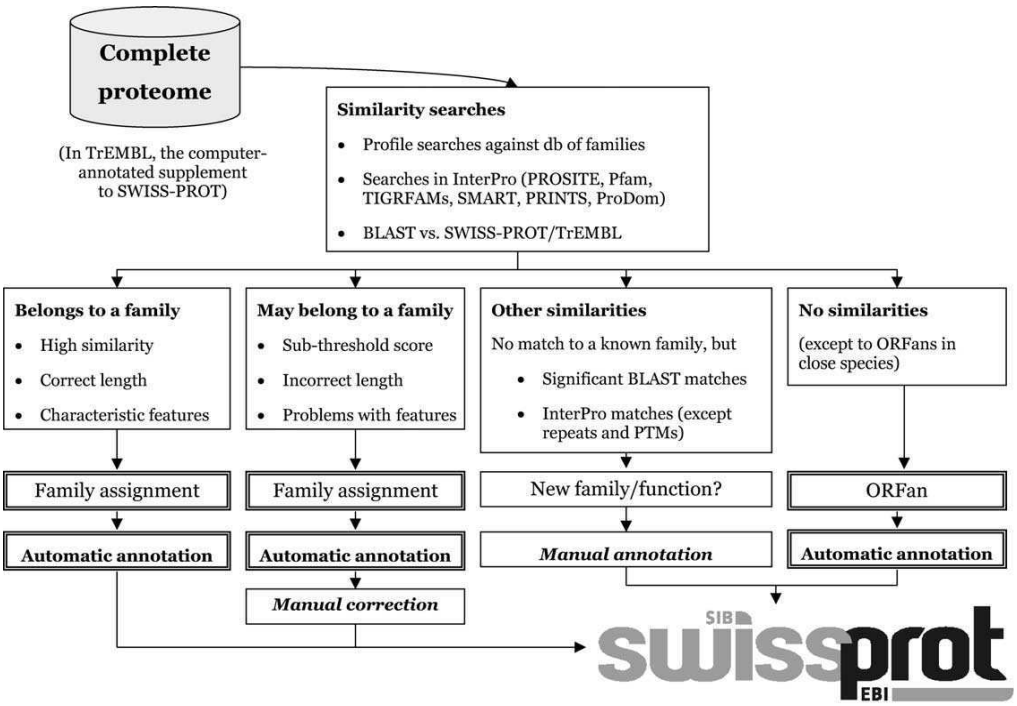
FIG. 4.2 – Architecture de la plate-forme d'exploration des génomes procaryotes *Genostar*

Un module (*e.g. GenoAnnot*) de la plate-forme de génomique exploratoire *Genostar*, s'appuie sur deux notions essentielles du service représentation de connaissances *GenoCore*.

Premièrement, la base de connaissances décrit les entités et les concepts manipulés par ce module (par exemple la notion de séquence ou encore de zone codante sur cette séquence sont des notions représentées dans le module *GenoAnnot*). La représentation de ces connaissances est donc assurée par *GenoCore* qui lui même utilise et étend une librairie nommée *AROM* (Allier Relations et Objets pour Modéliser). Cette librairie permet de représenter des entités décrites dans la base de connaissances par des objets informatiques qu'il devient alors possible de manipuler via une interface de programmation (API).

Deuxièmement, les tâches sont le moyen privilégié de manipuler (c'est-à-dire de lire, créer, modifier ou supprimer) ces connaissances. La représentation et l'exécution de ces tâches et assurée par *GenoCore* au moyen d'une librairie nommée *AROMTasks*. Une tâche est donc un programme qui a pour but de manipuler les objets de la base de connaissances. Finalement, lorsque le module est lancé, il charge la base de connaissances et les tâches qui lui sont associées (étapes 1 et 2).

Les librairies *AROM* et *AROMTasks* traduisent ces descriptions textuelles en des objets Java dans la machine virtuelle. Lorsqu'une tâche est démarrée, elle exécute du code qui lit, écrit ou modifie le contenu de la base de connaissances (étape 3). Ce code réalise cela en utilisant l'API de la librairie *AROM*. Le code exécuté par la tâche est un code écrit en langage Java qui est interprété à chaque exécution. Ce code peut cependant faire appel à des librairies dites *externes* (étape 4) ; c'est à dire des librairies écrites par exemple en C et compilées pour une architecture donnée (Linux, Solaris ou MacOS par exemple). Cette figure est extraite de `http://www.genostar.org/members/developpeur/overview.html`.



The workflow of the automatic annotation process in HAMAP . Complete microbial proteomes in TrEMBL are searched against the family prole collection, and conden t matches are automatically annotated based on the corresponding family rule. Only in the case where no warning is produced may the entry enter SWISS-PROT directly without manual verication. Entries with no similarities found by BLAST and InterPro family and domain scanning are the scope of 'ORFans' automatic annotation. The remaining entries, which are still the vast majority of entries in any given proteome, are subject to classical manual annotation, but may serve to 'seed' new family rules. Other tasks performed in HAMAP (with diverse degrees of automation) include redundancy removal, cleanups performed in TrEMBL, generation and update of family rules, sequence corrections and updates of SWISS-PROT entries.

FIG. 4.3 – Cette figure a été extraite de la publication *HAMAP* [Gattiker *et al.*, 2003].

Chapitre 5

Annotation *in silico* de génomes procaryotes

5.1 *Article I* : « L'annotation *in silico* des séquences génomiques »

C. Médigue, S. Bocs, L. Labarre, C. Mathé, D. Vallenet (2002)
Medecine/Sciences, 18, 237-250.

Si les expériences *humides* permettent d'apporter des résultats importants dans l'organisation et le fonctionnement des organismes bactériens, leur coût en temps et en moyens n'en fait pas un moyen d'analyse systématique de l'ensemble des génomes séquencés. Cependant, la prédiction des fonctions biologiques, à partir de l'ensemble des connaissances biologiques disponibles (séquences, données expérimentales), même parcellaires, est absolument nécessaire à la valorisation des efforts de séquençage (Fig. 5.1 p. 166). La technique des puces à ADN permet de mesurer et de visualiser très rapidement l'expression des gènes dans différentes conditions et ceci à l'échelle d'un génome complet. Nous espérons donc qu'elle aidera à rétablir l'équilibre entre la quantité de données prédites et la quantité de données expérimentales. Pour l'instant l'analyse bioinformatique des résultats de cette technique de biologie moléculaire est en cours de validation. Nous pourrions ainsi résumer le premier volet de ce manuscrit en reprenant les trois objectifs biologiques principaux auxquels fait référence l'annotation d'un génome procaryotes, passerelle entre la séquence génomique et la biologie de l'organisme [Stein, 2001] :

1. L'annotation syntaxique concerne l'identification de zones d'intérêt sur la séquence. Il s'agit typiquement de la recherche des zones codant potentiellement pour des protéines ou des ARNt, de la recherche de signaux de régulation de l'expression génétique et, d'une manière générale, de la localisation de motifs lexicaux ou structuraux caractérisés. A l'heure actuelle, si l'annotation syntaxique de gènes codant les protéines des génomes procaryotes est parfois délicate, c'est en partie dû aux problèmes d'identification des signaux de régulation. De même, l'identification des petits ARN fonctionnels reste un problème pertinent.

2. L'annotation fonctionnelle concerne l'attribution d'une (ou plusieurs) fonction(s) biologique(s) aux objets détectés au niveau précèdent. L'exemple typique est l'attribution d'un rôle fonctionnel aux produits protéiques des gènes ou la caractérisation fonctionnelle d'une séquence opératrice. Lorsqu'il n'existe pas de données expérimentales associées à une séquence polypeptidique (le produit d'un gène), nous avons vu que la stratégie classique consiste à effectuer un criblage des bases de séquences afin d'identifier des séquences fortement similaires et à attribuer, par analogie, leur(s) fonction(s) à la séquence requête. Les résultats d'une telle stratégie sont des hypothèses de travail qu'il convient, dans l'idéal, de valider expérimentalement. Réalisée automatiquement, cette stratégie d'assignation de fonctions présente de nombreuses limites. Par exemple, il est nécessaire d'évaluer au cas par cas la pertinence de la similarité entre les séquences comparées. Aussi nous avons vu que cette stratégie est totalement dépendante de la qualité des données présentes dans les bases de séquences publiques utilisées lors du criblage (problème de propagation des erreurs). Enfin, les relations entre les entités manipulées ne sont pas exploitées. Ainsi, on n'exploite encore que trop peu ou pas systématiquement le fait que des enzymes (protéines ayant la fonction de catalyser des transformations chimiques) intervenant dans une même voie métabolique (ensemble de réactions chimiques couplées) tendent à être groupés en opérons (groupe de gènes co-transcrits et donc co-localisés sur le chromosome). L'annotation fonctionnelle consiste aussi à définir de couples d'orthologues et de couples de paralogues.
3. Ainsi, l'annotation relationnelle concerne 1'identification des relations existant entre les objets caractérisés (individuellement) aux deux niveaux précédents. Ces relations sont de natures diverses. Il peut s'agir par exemple de leur implication dans un processus cellulaire commun (participation à une même voie métabolique, à une même voie de transport), ou d'une interaction physique (interaction protéine-protéine). Les informations qui doivent être manipulées à ce niveau d'annotation (opérons, régulons, synténies, graphes représentant des chemins réactionnels ou des assemblages moléculaires) sont plus complexes que les seules données de séquences et réclament donc un traitement particulier. Les objets manipulés et les relations qu'ils entretiennent présentent généralement un plus haut degré d'abstraction et de structuration (un graphe décrivant un réseau métabolique). Il se pose alors deux problèmes majeurs : d'une part, le problème de leur représentation formelle, c'est-à-dire leur modélisation, et d'autre part le problème de leur instanciation. Concernant l'aspect modélisation, plusieurs initiatives ont déjà vu le jour avec l'objectif de représenter ces informations nouvelles : *EcoCyc*[1] ou *KEGG*[2] pour les données métaboliques, et *RegulonDB*[3] pour les données d'opérons. Pour l'instant, ces efforts ne sont malheureusement que peu ou pas concertés.

Cette publication donne une vue d'ensemble de l'annotation des séquences génomiques procaryotes et eucaryotes. Elle décrits les principales stratégies d'analyse informatique mises en œuvre à

[1] `http://ecocyc.panbio.com`
[2] `http://www.genome.ad.jp/kegg/`
[3] `http://www.cifn.unam.mx/Computational-Genomics/regulondb/`

chaque niveau d'annotation : syntaxique (prédiction d'objets biologiques), fonctionnelle (recherche de similitudes, d'orthologues, de paralogues), et relationnelle (identification de synténies, d'opérons, de régulons).

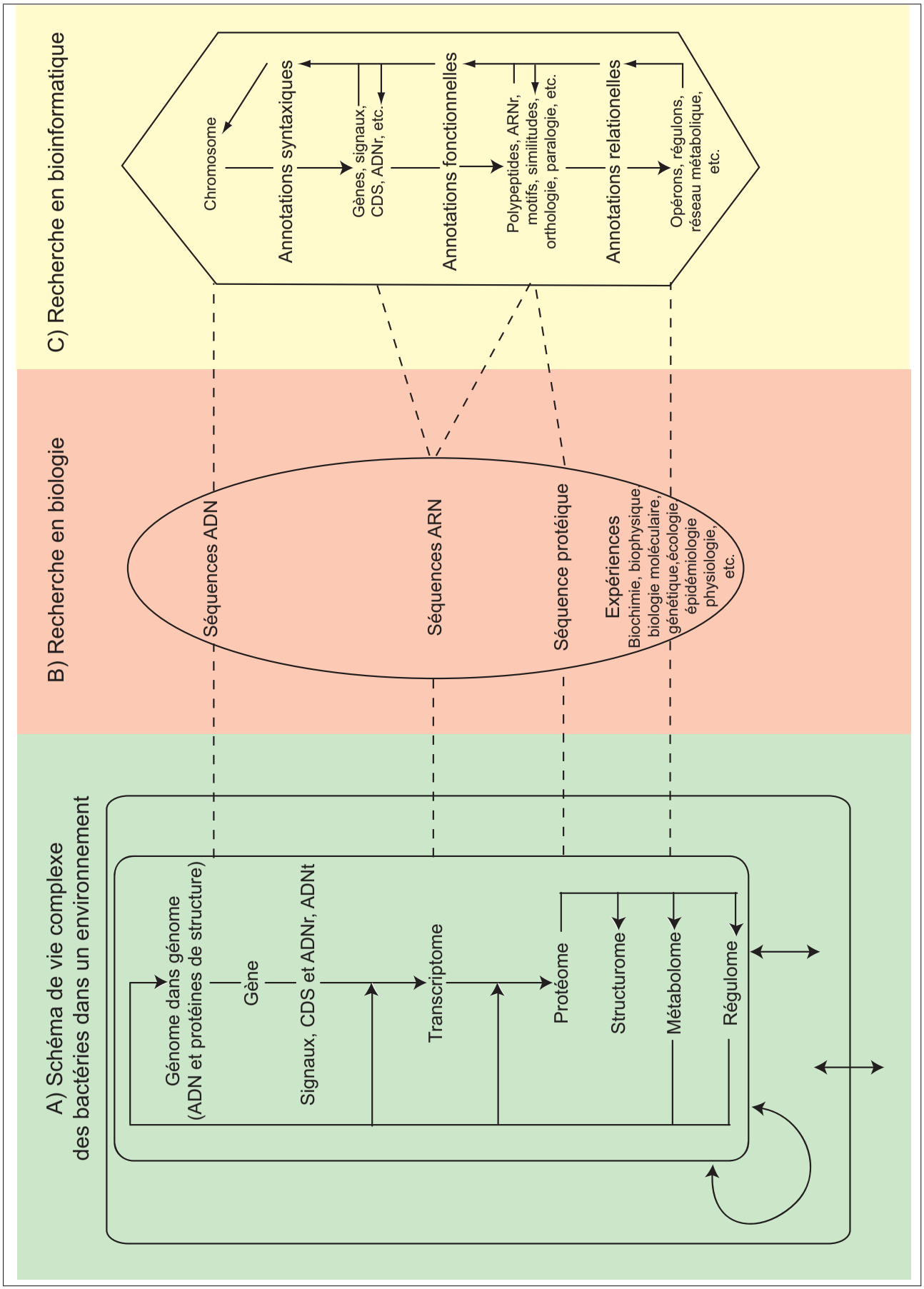


FIG. 5.1 – Relations complexes entre les recherches en biologie et en bioinformatique appliquée au monde procaryote

FIG. 5.1 – Relations complexes entre les recherches en biologie et en bioinformatique appliquée au monde procaryote

A) Schéma de vie complexe des procaryotes dans un environnement : le cadre interne symbolise les bactéries et le cadre externe symbolise l'environnement (une cellule hôte). Les flèches unidirectionnelles représentent les processus de synthèse et de régulation (réplication, transcription, traduction). Les flèches bidirectionnelles représentent des voies d'échanges entre les bactéries, entre les bactéries et leur environnement (communication par import-export de signaux, nutrition-excrétion de métabolites, transferts génétiques).

B) La recherche en biologie permet d'une part de séquencer les polymères nucléiques et protéiques ; et d'autre part de comprendre les processus bactériens par des expériences portant sur la fonction, la structure, les interactions de ces molécules et de ces organismes dans différentes conditions.

C) La recherche en bioinformatique utilise les données de la biologie pour prédire les processus bactériens. Ces prédictions se font par affinage progressif. En effet, vu la complexité des processus biologiques, il est illusoire de vouloir créer un outil presse bouton et tout en un. A chaque étape de l'annotation, les données servent
(1) de point de départ pour l'étape suivante
(2) de résultats *in silico* validant ou non les données de l'étape précédente, d'une part pour réajuster les paramètres des méthodes et d'autre part pour réévaluer les résultats
Par exemple, les données d'annotation syntaxique servent
(1) de point de départ pour l'annotation fonctionnelle
(2) à apprendre les paramètres des méthodes d'annotation syntaxique et à corriger les erreurs de séquençage dans le chromosome
De même, les données d'annotation fonctionnelle servent à l'annotation relationnelle mais aussi, aident à réajuster le codon d'initiation des CDS issues de l'annotation syntaxique. De même, les résultats de l'annotation relationnelle permettent de réévaluer la fonction d'un polypeptide qui était au départ inconnue, etc. Finalement les traits pointillés indiquent des relations entre le monde procaryote, les recherches en biologie et en bioinformatique. Notamment les prédictions bioinformatiques peuvent d'une part ouvrir de nouveaux axes de recherche sur des problématiques biologiques et d'autre part doivent être, autant que possible, confirmées par des expériences biologiques. Ainsi biologistes et bioinformaticiens s'évertuent à reproduire les processus biologiques dans le but de les comprendre. Les expériences biologiques permettent par exemple, de couper de l'ADN grâce aux enzymes de restrictions, de l'amplifier grâce aux ADN polymérases, de définir et transférer des gènes (transformation et recombinaison), d'isoler des ARN, de purifier des complexes protéiques fonctionnels, de réaliser des réactions biochimiques. De même, les méthodes d'analyse bioinformatiques, permettent de copier, définir, transcrire, traduire les gènes, de reconnaître les séquences similaires, de reconstruire les réseaux cellulaires, etc.

Deuxième partie

Développements d'outils pour annoter des génomes procaryotes

L'Atelier de Génomique Comparative a pour premier objectif d'explorer des données d'annotation bactérienne par génomique comparative. Pour cela, nous avons besoin en premier lieu de jeux de CDS correspondant aux génomes bactériens complets. Nous avons vu que les différentes stratégies d'annotation des centres de séquençage étaient en partie responsables de l'hétérogénéité d'annotation entre les génomes (voir p. 65 et p. 152). Ces stratégies comportent généralement une étape de validation manuelle par des annotateurs, générant une hétérogénéité d'annotation à l'intérieur d'un même génome (voir p. 311). En génomique comparative, on s'intéresse d'abord aux gènes communs entre les génomes puis aux gènes uniques à un génome. Si l'on compare des annotations hétérogènes, on ne sera pas capable de déterminer si l'absence d'un gène est un fait biologique ou simplement dû à une erreur d'annotation. C'est pourquoi, le premier objectif que nous nous sommes fixés est de mettre en place un processus qui permette de vérifier la qualité des jeux d'annotations disponibles dans la banque INSD, d'homogénéiser si besoin est ces annotations, mais aussi d'annoter un nouveau génome dans le cas où la séquence serait disponible sans annotation.

Selon les connaissances disponibles pour explorer un génome procaryote, nous sommes généralement amenés à l'annoter ou à le réannoter. Brièvement, nous nous servons d'abord de modules de la plate-forme *Genostar* pour étudier l'usage des codons synonymes et prédire des gènes et d'autres objets génomiques (Tab. 5.2 B p. 172). Puis, nous utilisons la plate-forme Biofacet et d'autres programmes comme InterProScan pour prédire d'autres objets biologiques comme l'alignement de séquences protéiques similaires et l'alignement de motifs protéiques. Tous les résultats de ces prédictions automatiques syntaxiques et fonctionnelles sont stockées dans une base de données relationnelle : *Prokaryotic Genome DataBase* (PkGDB). Ensuite, des annotateurs valident manuellement ces prédictions biologiques en utilisant l'interface web cartographique, *Magnifying Genomes* (*MaGe*), développée au dessus de PkGDB. Enfin, pour la phase finale d'exploration des connaissances, les deux plates-formes, *Genostar* et PkGDB–*MaGe*, sont utilisées en fonction des questions à résoudre. Par exemple, il est très intéressant de pouvoir explorer simultanément les groupes de gènes atypiques, les groupes de synténies et les voies métaboliques.

Tout au long du processus de (ré)annotation présenté dans la *deuxième partie*, consacrée aux développements d'outils, la base PkGDB permet de centraliser les données primaires (données des banques) et secondaires (résultats d'analyse ; Tab. 5.2 A p. 172 et voir le *sixième chapitre* p. 175). Dans l'état de l'art sur les méthodes de prédiction de gènes, nous avons compris que les meilleures méthodes étaient celles qui dépendaient d'un modèle probabiliste de séquences d'ADN. La précision de telles méthodes dépend avant tout du modèle choisi et de la qualité de la phase d'apprentissage des paramètres du modèle. Le non-codant doit être modélisé par des séquences natives du génome plutôt que par une séquence aléatoire. L'hétérogénéité des CDS, même si elle n'est pas directement modélisée, doit être prise en compte d'une manière ou d'une autre. Le *septième chapitre* est donc consacré à une stratégie de construction de matrices de probabilités de transition de chaînes de Markov pour chaque classe de gènes du génome, définie en fonction de l'usage des codons synonymes, et pour un jeu de séquences non-codantes (*AMIGene Matrices* (*AMIMat*) ; Tab. 5.2 C p. 172 et voir p. 197).

A

Ressources	Catégorie	Contenu	Référence
PkGDB	Base de données multigénome	SGBDR MySQL	agc@genoscope.cns.fr
RefSeq	Collection non-redondante de séquences de génomes de référence	SGBDR Oracle SQL (XML)	
EcoGene	Banque génomique d'*E. coli*	Fichiers à plats	
GenoList	Bases génomiques Subtilist, Tuberculist	SGBDR Sybase SQL	Tableau 2.1
SWALL	Banque de séquences protéiques annotées	Fichiers à plats (XML)	
InterPro	Base de motifs protéiques	SGBDR Oracle SQL (XML)	
PEC	Base des gènes essentiels d'*E. coli*	SGBD	http://www.shigen.nig.ac.jp/ecoli/pec/
GenProtec	Banques de séquences protéiques annotées d'*E. coli*	Fichiers à plats	
COG	Banque de génomique comparative	Fichiers à plats	Tableau 2.1
Enzyme	Banque des numéros de la commission Enzyme (EC)	Fichiers à plats (ASN.1)	
BSORF	Base des gènes essentiels de *B. subtilis*	SGBD dédié Fichiers à plats (XML) et système	Kobayashi-2003 http://bacillus.genome.ad.jp/
KEGG	Serveur de voies métaboliques	d'interrogation DBGET - LinkDB	
RegulonDB	Base de régulations cellulaires	SGBDR	Tableau 2.1
BIND	Base d'interactions moléculaires	SGBD dédié orienté objets (UML et XML)	

B

Plates-formes	Catégorie	Contenu	Référence
Biofacet (LASSAP)	Génomique comparative	L'architecture de cette plate-forme (encore appelée "package") est constituée de trois composants: un SGBD, un moteur de comparaison de séquences et un système pour l'analyse des résultats	Figure 4.1
R	Analyse statistique de données	Ce package intègre des méthodes de calcul statistiques et de représentation graphique	R-2003
Statistica	Analyse statistique de données	Ce logiciel intègre des méthodes d'analyse statistique inférentielle	http://www.statsoft.com/
Imagene	Environnement pour annoter et analyser les séquences	Système de représentation de connaissances factuelles et méthodologiques par objets (SRCO) et visualisation graphique dynamique de ces connaissances	Médigue-1999-a
Genostar	Environnement pour explorer les génomes	Les modules comme GenoAnnot, GenoBool et GenoLink reposent sur un SRCO et sur une visualisation graphique (sucesseur d'Imagene)	Figure 4.1
MAGE	Système d'annotation des génomes procaryotes	Cette interface graphique dynamique d'annotation repose sur PkGDB en particulier sur les synténies	Vallenet en préparation

FIG. 5.2 – Outils de l'Atelier de Génomique Comparative : A) Ressources B) Plates-formes

C

Méthodes	Catégorie	Contenu	Référence
Analyse_seq	Réplicon	Pourcentage en G+C, GCskew, GCskew cumulatif	http://ludwig-sun2.unil.ch/~bsondere/tp1/
Oriloc		GCskew cumulatif pour déterminer l'origine et le terminus	Franck-2000
Nosferatu		Recherche de répétitions longues (20 pb) non exactes par alignement local (Smith et Waterman)	Tableau 3.1
tRNAscan-SE	ARNt	Recherche de motifs et de structures secondaires	Tableau 3.1
Petrin	Terminateur	Recherche de structures tiges - boucles	Tableau 3.1
FindrRNA	ARNr	Recherche de similitudes (Blast2n de Biofacet)	alajus@genoscope.cns.fr
RBSfinder	RBS	Découverte et recherche de RBS	Tableau 3.1
Spat	motif	Recherche de motifs (Imagene)	Tableau 3.1
Spoc	CDS	Recherche de CDS par signal (Imagene)	Alain.Viari@inrialpes.fr
Prokov		Modules pour la prédiction de gènes procaryotes basés sur les chaînes de Markov	Tableau 3.3
AFCcodons		Analyse factorielle de correspondances (RSCU)	http://chlora.infobiogen.fr:1234/~chiapell/afc/afc.html
Ether		Partitionnement automatique (itération des nuées dynamiques Genostar)	Francois.Rechenmann@inria.fr
AMIMat		Phase d'apprentissage pour la prédiction de gènes utilise un programme comme Prokov	http://www.genoscope.cns.fr/agc/tools/amigene/
AMIGene		Phase de reconnaissance de la prédiction de gènes utilise un programme comme Prokov	Bocs-2003
CompAnnot	CDS et polypeptide	Comparaison de deux jeux de CDS d'un même chromosome	Bocs-2002
Seg		Masquage des régions de basse complexité compositionnelle dans les séquences protéiques	Wootton-1993
Xnu		Masquage des régions contenant des répétitions de courte périodicité dans les séquences protéiques	Claverie-1993
SWAN		CDS et polypeptide	Bocs-2002
CodonW		Calcul d'indices et AFC	Peden-1999 http://www.molbiol.ox.ac.uk/cu/
ProFED	Réplicon et CDS	Prédiction différents type de décalages du cadre de lecture qui utilise un programme comme Prokov	Médigue-1999-b
InterProScan	Polypeptide et protéome	Stratégie qui combine différentes méthodes de reconnaissance de signatures protéiques, natives aux banques intégrées dans InterPro	Zdobnov-2001
TMHMM		HMM pour la prédiction d'hélices alpha transmembranaires dans les protéines	Krogh-2001
SignalP		Combinaison d'une recherche de site de clivage, et de réseaux de neurones et d'un HMM pour la reconnaissance de peptide signal	Nielsen-1997
COGnitor		Assignation d'un polypeptide à un COG sur la base de "best hits" avec de multiples génomes spécifiques	Tableau 2.1
PRIAM		RPS-BLAST permet de rechercher des similitudes avec des PRofils pour l'Identification Automatique du Métabolisme et d'assigner un ou plusieurs numéros EC à un polyptide	Claudel-Renard-2003
Syntonizer	Protéome	Prédiction des groupes de synténie procaryote (modélisation sous forme de graphes). L'utilisateur peut choisir les génomes et définir les relations de correspondance et de colocalisation entre les gènes.	Labarre en préparation
PkGDB_tool	SGBD	Programmes de chargement de différents types d'annotations et de prédictions dans PkGDB et d'extraction des ces annotations dans différents formats	AGC en préparation

FIG. 5.2 – Outils de l'Atelier de Génomique Comparative : C) Méthodes)

Par ailleurs, nous ne nous sommes pas contentés d'une simple phase de reconnaissance de séquences d'ADN sous le modèle des chaînes de Markov. Notre expérience d'annotateur nous a permis d'enchaîner, en plus, une phase de filtrage des CDS détectées lors de la phase de reconnaissance. En effet, nous avons développé une heuristique pour filtrer les CDS les plus probables parmi celles détectées en combinant plusieurs critères. Le *huitième chapitre* explique la méthode de prédiction de gènes procaryotes (*Annotation of MIcrobial Genes* (*AMIGene*) ; Tab. 5.2 C p. 172 et voir p. 237). Le *neuvième chapitre* décrit le processus de réannotation d'un chromosome procaryote qui consiste à comparer le jeu de CDS annotées dans les banques au jeu de CDS prédites par notre heuristique, et à attribuer un statut de réannotation à certaines des CDS uniques en combinant plusieurs critères (Tab. 5.2 C p. 172 et voir p. 275).

Enfin, la *troisième partie* de ce manuscrit est dédiée à l'analyse de prédictions biologiques issus du processus de (ré)annotation appliquées à des génomes procaryotes complets. Le *dixième chapitre* et le *onzième chapitre* présentent respectivement, des résultats de réannotation de chromosomes, et des résultats d'annotation et d'exploration de chromosomes d'entérobactéries.

Chapitre 6

Base multigénomes PkGDB

Pour réaliser des comparaisons entre les génomes, les données d'annotations publiques et les données produites par les méthodes bioinformatiques doivent être intégrées dans un SGBD. Durant ma thèse, j'ai participé à la conception du modèle, au chargement et à la mise à jour des données de la base multi-génomes de l'Atelier de Génomique Comparative : *Prokaryotic Genome DataBase* (PkGDB) [Lefebvre, 1999, Labarre, 2000, Bentz, 2000, Devine, 2001, Vallenet, 2002, Jackson, 2002]. Cette base permet de modéliser les données des génomes procaryotes complets. Comme toute base de données, elle est constituée de trois parties : (i) le SGBD (SGBDR MySQL), (ii) le modèle de données (Fig. 6.1 p. 176) et (iii) les données primaires (les ressources ; Tab. 5.2 A p. 172) et secondaires (les résultats de méthodes ; Tab. 5.2 C p. 172).

Le modèle PkGDB est générique car il peut être utilisé pour un projet d'annotation d'un nouveau génome (*e.g.* instance *Acineto*DB), pour un projet de réannotation des génomes complets publiés (*e.g.* instance NeisseriaDB) ou pour un projet d'exploration selon une thématique particulière (voir p. 195). Par exemple l'instance *Entero*DB, permet d'explorer les annotations de génomes d'entérobactéries pathogènes (*e.g. Escherichia coli* O157 :H7 EDL933, *S. enterica* serovar Typhimurium LT2, *P. luminescens*, *Y. pestis* CO92) en référence aux annotations d'un génome modèle « proche » (*e.g.* entérobactérie commensale *E. coli* K-12). Une instance de base est générée grâce à MySQL qui permet d'intégrer des données appropriées dans le modèle PkGDB. Par exemple, dans le cas de l'instance NeisseriaDB, des données nécessaires à la réannotation et à l'exploration des génomes des *Neisseria* alimentent la base : (i) les annotations des réplicons contenues dans les fichiers des banques INSD et des prédictions automatiques de l'Atelier de Génomique Comparative, (ii) des résultats de comparaison de deux jeux de CDS d'un même réplicon (statut de réannotation) et (iii) des résultats de comparaison des jeux de CDS entre des génomes bactériens (prédiction de groupes de synténie).

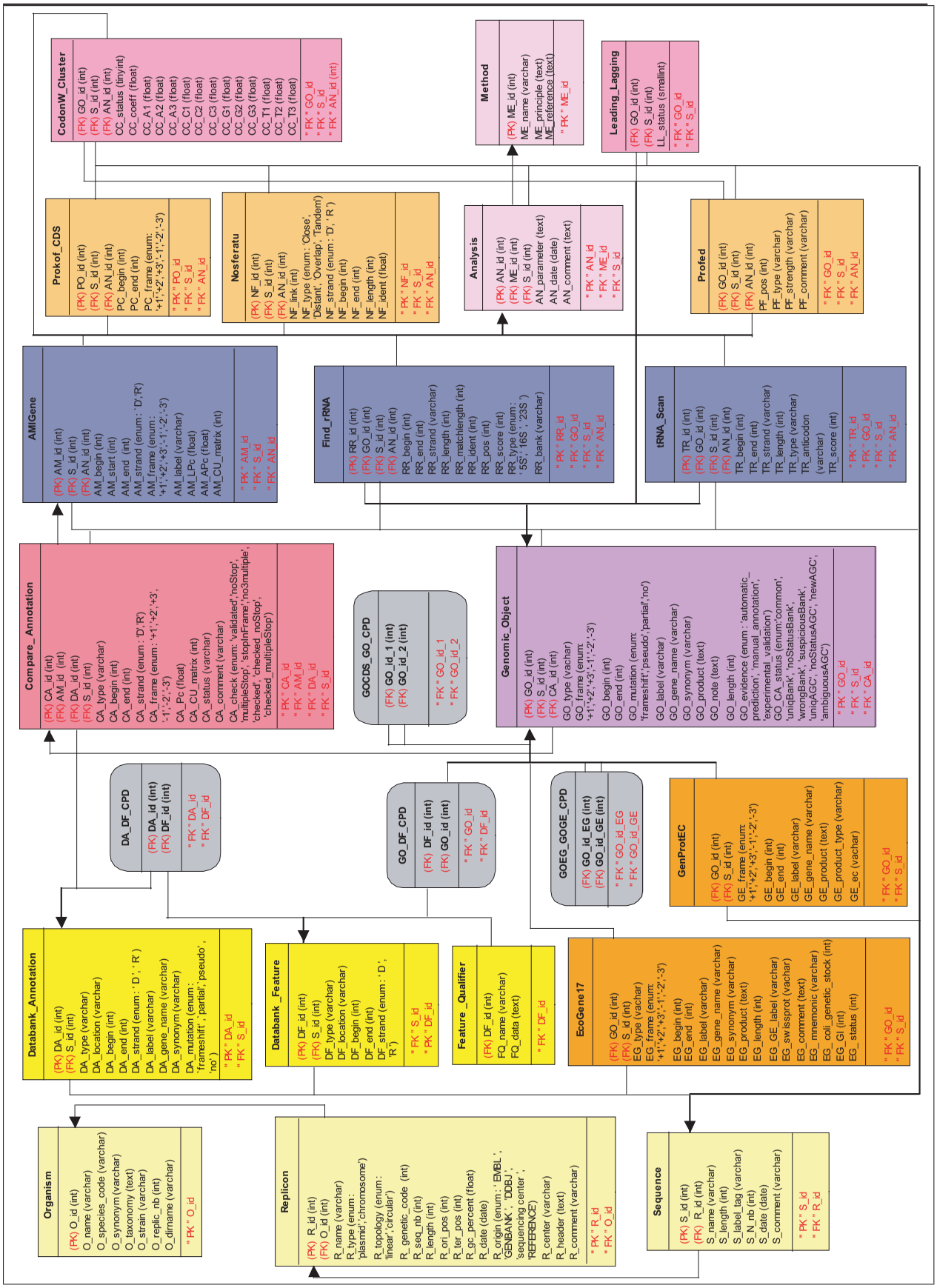


FIG. 6.1 – A) *Prokaryotic Genome DataBase* (annotation des séquences nucléiques)
Diagramme entités–associations selon un formalisme dérivé d'UML.

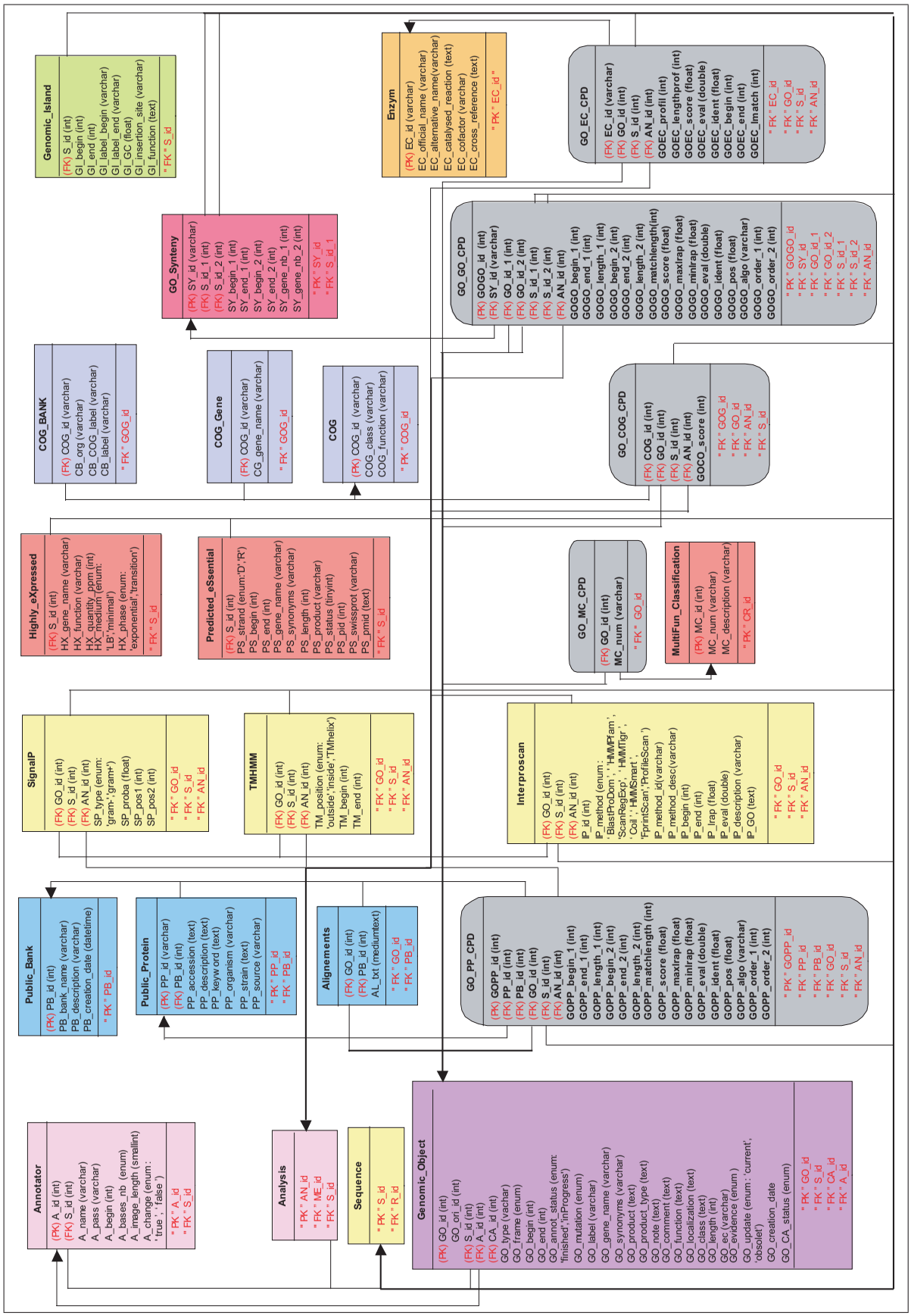


FIG. 6.1 – B) PkGDB (annotation des séquences protéiques)

6.1 Points clés de la structure logique générique PkGDB

Le schéma conceptuel actuel de la base de données est présentée dans la figure 6.1 p. 176. Il est issu de réflexions communes qui se sont dégagées depuis plusieurs années au cours des discussions animées des réunions de l'Atelier de Génomique Comparative. Il est régulièrement remis en question, remanié selon les besoins et éprouvé (aller-retour entre la théorie et la pratique). Les informations sont modélisées dans plusieurs tables : certaines servent à décrire un type d'entités biologiques et d'autres permettent d'établir des correspondances $n \leftrightarrow m$ entre deux types d'entités biologiques ($n, m \geq 1$; le nom de ces tables se terminent alors par *_CPD*). Dans la suite de ce chapitre, les tables seront notées entre crochets et les champs commencent par un code à deux ou quatre lettres majuscules suivi d'un tiret bas. Les champs se terminant par *_id* sont des clefs primaires ou étrangères. Elles sont indispensables pour relier les tables entre elles (jointures) et doivent être indexées pour optimiser le temps d'exécution des requêtes. Les clés essentielles de ce modèle sont : (i) l'identifiant d'une séquence nucléique (S_id) de la table [Sequence], (ii) l'identifiant d'un objet génomique (GO_id) de la table [Genomic_object] et (iii) l'identifiant d'une méthode d'analyse (AN_id) de la table [Analysis].

La table [Organism] contient le nom et la description de tous les organismes présents dans la base. Le génome d'un organisme peut être constitué de plusieurs unités de réplications (chromosomes, plasmides) stockées dans la table [Replicon]. La table [Sequence] permet d'associer plusieurs séquences à un même réplicon, par exemple lors de la mise à jour d'une séquence (nouvelle version du chromosome) ou lorsque que le réplicon est en cours de séquençage (contigs).

La table centrale de PkGDB, à laquelle s'adossent des interfaces graphiques (*e.g. MaGe* ; Tab. 5.2 B p. 172) est la *table des objets génomiques* [Genomic_Object]. Elle regroupe les objets d'intérêt biologique localisés sur les réplicons tels que des signaux de régulation et des séquences fonctionnelles (annexe F p. 411). Ces objets sont de types différents : *RBS*, *terminator*, *CDS*, *tRNA*, *rRNA*, etc. Les champs GO_begin, GO_end et GO_frame localisent l'objet sur le réplicon. Le champ GO_CA_status permet à la fois de distinguer l'origine de ces objets (*i.e.* banques publiques ou analyse *in silico* de l'Atelier de Génomique Comparative) et d'attribuer un statut de réannotation (voir p. 275). Une description biologique de l'objet est contenue dans les champs GO_product et GO_note. Le champ GO_mutation s'applique aux gènes codant des polypeptides : il indique la présence ou non d'une mutation dans la CDS et le type de mutation (*e.g.* décalage du cadre de lecture, codon de terminaison en phase).

Un objet génomique peut être issu de, ou caractérisé par, une ou plusieurs méthodes dont les résultats sont stockés dans une table portant généralement le nom de la méthode. Par exemple, une CDS peut être issue d'*AMIGene* ([AMIGene]) et contenir un décalage du cadre de lecture détecté par *ProFED* ([ProFED]). Toute analyse est identifiée (AN_id) afin de stocker les paramètres d'exécution (AN_parameter) et la date d'exécution (AN_date). De plus, une analyse renvoie à une méthode (ME_id). La table [Method] donne le nom de la méthode (ME_name), son principe (ME_principle) et sa référence (ME_reference).

Nous allons maintenant présenter la structure de PkGDB sous l'angle de la réannotation, car en pratique il est difficile de présenter la structure en faisant complètement abstraction de son instanciation, même si, en théorie, il est essentiel de ne pas mélanger ces deux concepts. Nous ne réannotons pas les ARNt et les ARNr (lorsque les annotations existent, nous faisons confiance aux banques).

6.2 Intégration des annotations de génomes complets de banques nucléiques

Les données à traiter sont contenues dans des fichiers à plat *RefSeq* (format GenBank) relatifs aux génomes procaryotes entièrement séquencés et annotés, distribués par la collaboration INSD. Cependant, pour certains génomes, il existe des bases génomiques qui fournissent une annotation de meilleure qualité (*e.g. EcoGene*). Le fichier tabulé de la banque *EcoGene* peut être directement chargé dans la table [EcoGene17]. Nous avons choisi de ne pas partir des fichiers *RefSeq* tabulés (format ptt) car il y a une perte d'information importante : seules les CDS sans /pseudo et seuls les attributs Location, Strand, Length, PID, Gene, Synonym, Product sont présents. Par exemple, dans le cas de *Y. pestis* CO92, 4008 CDS sont présentes dans NC_003143.gbk contre seulement 3885 dans NC_003143.ptt. L'intégration des données au format GenBank implique des traitements informatiques qui sont toujours suivis d'une vérification manuelle, les données provenant d'une même source n'étant pas forcément homogènes. Un problème majeur est rencontré avec les identifiants utilisés pour les objets génomiques : par exemple, une CDS peut être identifiée par son nom de gène, son label ou son numéro d'accession de banque. Cette diversité dans la dénomination des objets doit être formalisée pour associer un identifiant unique à chaque objet. Cette tâche peut être rendue encore plus difficile lors de la mise à jour de données et nécessite une vérification.

6.2.1 De l'organisme à la séquence

Pour plus de clarté, appelons PkGDB_Tools le processus de remplissage et d'interrogation de PkGDB qui regroupe tout un ensemble de programmes. Chacune des tables [Organisme], [Replicon] et [Sequence] est alimentée grâce à un programme de PkGDB_Tools.

1. Lors de l'analyse d'un génome, nous caractérisons d'abord l'organisme (un O_id est alors généré). Le programme utilise uniquement les annotations (fichier.gbk), et plus particulièrement l'entête de ce fichier (Fig. 2.1 p. 63). Il récupère le nom complet de l'espèce (genre, espèce et parfois souche), la taxonomie et construit le code *HAMAP*[1] (généralement les trois premières lettres du nom de genre sont concaténées avec les deux premières du nom d'espèce). Il demande à l'utilisateur les synonymes du nom d'espèce[2], le nombre de réplicons et la souche.

[1] `http://www.expasy.org/sprot/hamap/bacteria.html`
[2] `http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/`

Pour différentes souches d'une même espèce, il est recommandé de vérifier le code *HAMAP* (*e.g.* SALTI et SALTY, ou ECOLI et ECO57).

2. Le programme utilise le fichier des annotations, le fichier du chromosome (fichier.fna) et le O_id généré au cours de l'étape précédente. Il récupère automatiquement la valeur d'un certain nombre d'attributs[3]. Il demande à l'utilisateur R_origin (*e.g.* GenBank ou EMBL-EBI), R_center[4] (centre de séquençage), R_seq_nb (si le réplicon est complet, le nombre de contig vaut 1). Il calcule la longueur du chromosome et vérifie qu'elle est identique à celle donnée dans le fichier (sinon R_length prend la valeur calculée). Il calcule le pourcentage moyen en G+C du réplicon. Face à la séquence d'un nouveau chromosome complet, l'origine et le terminus de réplication sont généralement les premiers objets génomiques que l'on cherche à définir. Les positions de l'origine et du terminus, définies par exemple à partir du programme *Oriloc* [Frank & Lobry, 2000], sont stockées respectivement dans les champs *R_ori_pos* et *R_ter_pos* de la table *Replicon* (Fig. 6.1 A p. 176). A terme, ils deviendront des objets génomiques de la table [Genomic_object] (une origine de réplication est proposée uniquement, dans le fichier des annotations d'*E. coli* O157:H7 EDL933).
3. Le programme utilise le fichier des annotations, le fichier du chromosome et le R_id généré au cours de l'étape précédente. Il récupère S_date (publication du génome), S_comment (*e.g. complete sequence*), S_label_tag (*e.g.* BH pour reconnaître ensuite le label des CDS annotées comme BH0001 de *B. halodurans*), S_name (NC_002570.fna). Il demande à l'utilisateur S_label_tag s'il ne l'a pas trouvé. Il calcule aussi le nombre de bases indéterminées dans la séquence.

6.2.2 Annotations originales des chromosomes des banques

Dans le cas général des fichiers au format GenBank, les tables [Databank_Feature] et [Feature_Qualifier] contiennent les annotations brutes originales. Les attributs (ou *qualifiers*) sont liés à l'objet génomique (ou *feature*) par son identifiant (DF_id). Il existe aussi des tables qui sont spécialement conçues pour recevoir des annotations de référence de banques ou de bases génomiques ; c'est le cas par exemple de la table [EcoGene17] qui contient les annotations de référence d'*E. coli* K-12 (Tab. 5.2 A p. 172). Un programme de PkGDB_Tools permet de remplir les tables [Feature_Qualifier] et [Databank_Feature] à partir du fichier des annotations. Chaque objet annoté sur la séquence du chromosome (*Feature Table*) est découpé entièrement en qualificatifs (FQ_name et FQ_data). Par exemple, pour *B. halodurans*, on a FQ_name = 'source' et FQ_data = '1..4202353' ; le second qualificatif de cet objet source est FQ_name = 'organism' et FQ_data = 'Bacillus halodurans'. Le premier attribut d'un objet correspond toujours à son type et à sa localisation, ce qui est nécessaire et suffisant pour créer un objet de [Databank_Feature]. L'insertion d'un tuple dans [Databank_Feature] permet de générer la clé primaire DF_id qui sert de clé étrangère à la table

[3] R_name (NC_002570), R_length, R_topology, R_date (date du fichier *RefSeq*), R_header, R_comment, R_genetic_code, R_type.

[4] `http://wit.integratedgenomics.com/GOLD/`

[Feature_Qualifier] (la clé primaire n'est pas nécessaire pour cette table). Si un *join* est présent dans la localisation d'un objet (*e.g.* join(1277103..1278062,1280022..1280114) ; Fig. 2.2 p. 67) alors DF_begin et DF_end correspondent respectivement, à la première position de début (1277103) et à la dernière position de fin (1280114). Pour les FQ_data, les sauts de lignes (\n) et les guillemets ("") sont supprimés. Si la séquence protéique était sur plusieurs lignes, elles sont concaténées sans laisser d'espace ; pour les autres valeurs sur plusieurs lignes, elles sont concaténées en laissant un espace. Pour les attributs qui n'ont pas de valeur comme /pseudo, le FQ_data prend la valeur NULL. Le /partial est amené à disparaître. La norme EMBL-EBI conseille d'utiliser les signes '>' et '<' dans la description de la localisation (*e.g.* <345..500) pour indiquer que la séquence est partielle[5].

6.2.3 Homogénéisations automatiques et manuelles des annotations

Même si le format des banques comporte un certain nombre de règles précises, les annotations ne sont pas homogènes d'un organisme à l'autre et varient parfois au sein d'un même fichier. En particulier, des problèmes majeurs sont rencontrés pour détecter des annotations de *frameshift* et extraire les labels, les noms symboliques de gènes et leurs synonymes (Fig. 2.1 p. 63 et voir p. 65).

La table [Databank_Annotation] contient uniquement les gènes de type *CDS*, *tRNA* et *rRNA*. Elle rassemble les informations nécessaires pour définir et identifier correctement ces gènes. Les champs DA_type et DA_mutation permettent de typer le gène (*e.g. CDS*). Les champs DA_location, DA_begin, DA_end, DA_strand permettent de le localiser, et les champs DA_label, DA_symbol et DA_synonym permettent de l'identifier. La table de correspondances [DA_DF_CPD] permet de fusionner les doublons, c'est-à-dire les objets qui ont la même position de fin biologique (*i.e.* à chaque objet de type *gene* correspond généralement un objet de type *CDS*, *tRNA* ou *rRNA*).

Un programme de PkGDB_Tools utilise les tables [Feature_Qualifier] et [Databank_Feature] pour remplir les tables [DA_DF_CPD] et [Databank_Annotation]. Ce programme permet de résoudre certains problèmes de compatibilité sémantique. La différence essentielle entre les formats GenBank et EMBL-EBI est que tout objet de type *gene* est normalement associé à un objet de type *CDS*, *mRNA*, *tRNA*, *rRNA*, *misc_RNA* ou *scRNA*. Ces couples ont la même orientation sur le chromosome (directe ou inverse) et sont généralement définis par les mêmes bornes : c'est d'après ces critères que nous déterminons que deux objets représentent la même entité.

Plus généralement, le programme récupère d'abord l'information dont il a besoin. L'ensemble des gènes et l'ensemble des CDS, mRNA, tRNA, rRNA contenus dans la table [Databank_Feature] sont stockés respectivement dans le tableau associatif Gene et dans le tableau associatif Feature. Les qualificatifs /gene, /label, /note, /pseudo, /partial de ces objets sont récupérés dans le tableau associatif Qualifier à partir de la table [Feature_Qualifier]. Ensuite, l'information est analysée. Pour chaque gène du tableau Gene, les objets du tableau Feature inclus dans le gène et les attributs du tableau Qualifier associés à ces objets sont analysés dans le but de renseigner [Databank_Annotation]. Cette étape permet de fusionner des objets et d'homogénéiser la valeur des attributs. Par exemple, dans

[5] `http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html`

le cas d'un couple gène–CDS, un seul objet de type *CDS* sera inséré dans [Databank_Annotation] en extrayant l'information nécessaire pour remplir les champs DA_ à partir de leurs qualificatifs. Enfin, un objet est inséré dans [Databank_Annotation]. Le DA_type peut prendre trois valeurs : *CDS*, *tRNA*, *rRNA* correspondant au DF_type de l'objet associé au gène.

Il existe au moins cinq façons différentes d'annoter un décalage du cadre de lecture dans une séquence codante (Fig. 2.2 p. 67) :

1. L'objet génomique de type *CDS* et/ou *gene* possède un champ /pseudo (exemple A ; Fig. 2.2 p. 67).
2. L'objet CDS et/ou gène est décrit avec le mot-clef *join* dans sa localisation (exemple A). L'objet est alors composé de plusieurs morceaux de séquences. Ces objets génomiques correspondent le plus souvent à des gènes avec un *frameshift* ou à des gènes d'ARN avec un intron de type II.
3. L'objet CDS contient dans la description du champ /note des mots-clefs indiquant la présence éventuelle d'un *frameshift* (*i.e.* 'frameshift', 'valid start', 'pseudogene', 'internal stop', 'point mutation', etc ; exemple D).
4. L'objet gène, non associé à un objet CDS, possède dans la description du champ /note un des mots-clefs cité précédemment. Les auteurs qui annote les *frameshifts* de cette façon considèrent que ces gènes ne sont pas fonctionnels (pseudogène) ; c'est pourquoi ils n'associent pas d'objet CDS (le gène n'est pas traduit *in vivo*).
5. L'objet gène est associé à deux CDS, par exemple, le gène *metC'* chez *Campylobacter jejuni* est associé à deux fragments de CDS *metC'*.

Dans PkGDB, ces objets sont considérés comme des CDS contenant un *frameshift*. Le champ DA_location de la table [Databank_Annotation] contient la localisation éventuellement corrigée dans le cas d'une CDS constituée de plusieurs fragments (construction du *join* adéquat). Une CDS peut donc avoir un DA_mutation à 'pseudo' avec ou sans *join* dans le DA_location. Le champ DA_mutation prend la valeur 'frameshift' uniquement lorsque le décalage du cadre de lecture est programmé pour réguler la traduction de la CDS (*e.g. prfB* chez *E. coli* K-12, *B. subtilis*, etc.). Le label d'une CDS devrait logiquement être indiqué dans la description du champ /label de l'objet CDS ou gène, mais ce champ n'est pas utilisé. Pour remplir le champ DA_label, le programme retrouve généralement le label dans le /note de l'objet gène. Cependant, pour une CDS sans nom (symbol du gène, *e.g. dnaA*), le label est décrit dans le champ /gene. Il arrive même, dans ce dernier cas, que le label ait été tout simplement perdu. Nous avons vu dans le deuxième chapitre de l'état de l'art qu'un nouveau qualificatif /locus_tag/ est en train d'être mis en place par la collaboration INSD pour résoudre ces problèmes de label. Pour remplir le champ DA_gene_name, le nom symbolique du gène se trouve généralement dans le champ /gene de l'objet CDS ou gène. Pour résoudre le problème des noms de gène synonymes listés, soit dans le /gene, soit dans le /note de l'objet CDS ou gène, la collaboration a décidé d'utiliser le vocabulaire contrôlé 'synonym' dans le /note du gène des fichiers *RefSeq*.

Actuellement, un DA_id est lié au maximum à quatre DF_id, dans le cas d'un *join*, ou d'un couple en double. Il existe des couples gène–CDS en double (erreur d'annotation). Nous les reconnaissons car ils ont une position de codon de terminaison identique. Dans ce cas, une seule CDS sera créée dans la table [Databank_Annotation] et sera associée à quatre DF_id grâce à la table de correspondances [DA_DF_CPD]. Le programme est exécuté une première fois. Tous les gènes sans objet associé sont répertoriés dans un fichier. L'utilisateur retrouve le type *CDS*, *tRNA*, *rRNA* de certains objets qu'il corrige. Les objets non corrigés ne passent pas dans [Databank_Annotation] mais seront récupérés plus tard lors de l'alimentation de [Genomic_Object]. Le programme est exécuté une seconde fois avec le fichier des types corrigés.

6.2.4 Corrections automatiques et manuelles des bornes des CDS

Un programme de PkGDB_Tools copie les objets de la table [Databank_Annotation] dans la table [Compare_Annotation] en vérifiant la cohérence de bornes des CDS à l'aide de la séquence chromosomique. Les principales étapes de ce programme sont (i) l'extraction des données de [Databank_Annotation], (ii) la fragmentation des CDS qui contiennent des *join*, (iii) la vérification de la cohérence des bornes des CDS et (iv) le chargement de ces nouvelles données dans la table [Compare_Annotation].

Si une CDS possède un *join* dans le DA_location alors elle sera fragmentée dans [Compare_Annotation] ; c'est-à-dire que pour chaque fragment du *join*, un objet de type *fCDS* est inséré dans [Compare_Annotation].

Toute CDS ou fCDS doit être définie de la position du premier nucléotide d'un codon (qui n'est pas forcément un codon d'initiation) à la position du troisième nucléotide du premier codon de terminaison rencontré en phase. Ainsi, un programme attribue le statut de cohérence des bornes des CDS '*validated*', '*no3multiple*', '*noStop*', '*stopInFrame*' ou '*multipleStop*' aux CDS et fCDS (CA_check). De plus, les positions des CDS dont le codon de terminaison est localisé en +3 ou -3 de la position indiquée sont corrigées automatiquement et ont un CA_check à '*validated*' (certains auteurs utilisent la convention d'exclure le codon de terminaison des bornes des CDS). Dans le cas d'une CDS de [Databank_Annotation], avec un *join*, deux fCDS sont insérées dans [Compare_Annotation].

Dans la mesure du possible, les bornes des CDS anormales (CA_check != '*validated*') sont corrigées manuellement grâce à une interface cartographique dédiée à cette tâche (CompAnnotViewer). Cette étape de correction manuelle des CDS nécessite d'avoir construit au préalable des prématrices de transition à partir des CDS annotées extraites de [Compare_Annotation], d'avoir calculé les courbes de probabilités de codage et d'avoir chargé la table [Prokov_CDS] (voir p. 185 et p. 197). L'utilisateur est alors confronté à différents cas de figure que nous avons présentés dans la figure 6.2 p. 186.

La correction des bornes d'une CDS dont le statut est '*stopInFrame*' consiste souvent à la fragmenter en deux fCDS dont le statut passe à '*checked*' (dans ce cas, il y a mise à jour du tuple correspondant à la CDS originale, et insertion d'un tuple pour le second fragment). Si la CDS originale est dans le sens direct, il suffit alors de mettre à jour le CA_end au premier codon de ter-

minaison rencontré en phase en 3′. Lors de la création du fragment en 3′, c'est le codon d'initiation le plus en 3′ de ce dernier codon de terminaison qui est choisi. La fCDS en 3′ possède donc le codon de terminaison original mais pas le codon d'initiation.
Les cas où une séquence d'insertion (IS ; voir p. 46) est incluse dans une CDS méritent une attention particulière.

- L'IS fragmente souvent la CDS en deux parties : dans le *join* seuls les deux fragments de CDS doivent être raboutés (pour éviter de construire artificiellement une protéine chimérique).
- L'IS est aussi parfois insérée dans le sens inverse de la CDS (la définition du *join* suppose que tous les fragments sont dans le même sens, ce qui confirme que la CDS de l'IS n'a rien à faire dans le *join*).
- Des cas complexes peuvent se présenter comme un transposon composé (voir p. 46) inséré dans une CDS, ou comme des IS imbriquées, insérées dans une CDS.

Dans des cas rares et extrêmes, il arrive que l'utilisateur juge qu'aucune correction de borne ne puisse améliorer la définition d'une CDS anormale (pas de compromis satisfaisant entre les bornes proposées par *prokov_orf* et celles annotées par les auteurs). Il est alors autorisé à passer la CDS anormale en fCDS avec le statut '*checked_noStop*' (absence du codon de terminaison) ou avec le statut '*checked_multipleStop*' (présence d'au moins trois codons de terminaison[6] ; Fig. 6.2 p. 186).

Cette étape de corrections manuelles des bornes des CDS est suivie d'une vérification automatique d'une part de la complétion des corrections des CDS d'un génome des banques et d'autre part de la cohérence de ces corrections (il ne faut pas que la correction manuelle introduise une nouvelle erreur en voulant en corriger une). En pratique, cette étape correspond à la première étape de la phase de réconciliation des annotations des banques et des prédictions de l'Atelier de Génomique Comparative (voir p. 193). La complétion des corrections est vérifiée par une simple requête qui teste qu'il n'y a plus de CDS avec un CA_check à '*no3multiple*', '*noStop*', '*stopInFrame*'ou '*multipleStop*'. La cohérence des corrections est vérifiée en deux temps : (i) attribution d'un statut de cohérence des bornes des CDS comme précédemment et (ii) vérification de la cohérence de ce nouveau statut par rapport à la valeur actuelle de CA_check. Idéalement, tous les statuts sont à '*validated*' ce qui correspond à un CA_check à '*validated*' ou à '*checked*'. Exceptionnellement, on peut trouver les statuts '*noStop*' et '*multipleStop*' qui correspondent respectivement à un CA_check à '*checked_noStop*' et à '*checked_multipleStop*'. Dans tous les autres cas de figure, un problème de correction persiste. Par exemple il n'est pas cohérent qu'une CDS ait un statut à '*no3multiple*' et un CA_check à '*checked*'.

A ce stade (la cohérence des CDS a été vérifiée au moins trois fois), nous sommes prêts pour construire les matrices de transition nécessaires à la prédiction de CDS par *AMIGene* qui nous sert de témoin pour évaluer la justesse des annotations syntaxiques (voir p. 197 ; p. 185 et p. 193).
D'une manière générale, la cohérence de la valeur d'autres champs de la table [Compare_Annotation] devrait aussi être vérifiée. Par exemple, on pourrait vérifier que toutes les CDS qui ont un CA_check à '*checked_noStop*' ou à '*checked_multipleStop*' sont de type *fCDS* ou qu'il n'existe pas de CDS en

[6] S'il n'y a que deux codons de terminaison, il faut fragmenter la CDS.

doublon pour un même génome (deux CDS dans le même sens ayant la même position de codon de terminaison), etc.

6.3 Intégration des prédictions issues de résultats de méthodes

Une autre façon d'obtenir des annotations génomiques est d'utiliser des méthodes bioinformatique de prédiction (Tab. 5.2 C p. 172).

6.3.1 Prédictions syntaxiques

Les tables du modèle PkGDB dédiées à l'annotation syntaxique de la séquence d'un réplicon sont celles qui permettent de définir des objets génomiques de différents types. Les programmes tRNAscan-SE (resp. findrRNA) permettent d'identifier les gènes d'ARNt (resp. d'ARNr) stockés dans la table [tRNA_Scan] (resp. [Find_rRNA] ; Tab. 5.2 C p. 172). La table [AMIGene] contient les CDS prédites par la stratégie *AMIGene* (Tab. 5.2 C p. 172 et voir p. 197 et p. .237).

Les *frameshifts* sont détectés par le programme *ProFED* et stockés dans la table [ProFED] (Tab. 5.2 C p. 172). Différents types de répétitions longues sont recherchés sur le chromosome par le programme Nosferatu et les résultats sont stockés dans la table [Nosferatu] (Tab. 5.2 C p. 172). Actuellement, ces deux types d'objets ne sont pas intégrés dans la table [Genomic_object] ; ils sont cependant représentés dans l'interface cartographique de *MaGe* (Tab. 5.2 B p. 172). La table [Prokov_CDS] permet de définir toutes les CDS de longueur maximales supérieures à 60 pb.

Il est prévu d'intégrer d'autres méthodes d'annotation syntaxique dans le modèle PkGDB comme la recherche de RBS (RBSFinder) et de terminateurs rho indépendants (Petrin ; Tab. 5.2 C p. 172). Les recherches d'opérateurs, de promoteurs et de terminateurs dépendant du facteur rho ne sont pas effectuées car nous n'avons pas encore trouvé d'outils satisfaisants (le développement de nouveaux algorithmes n'est pas dans nos priorités).

6.3.2 Prédictions fonctionnelles

Les tables du modèle PkGDB dédiées à l'annotation fonctionnelle d'un réplicon permettent de caractériser les objets génomiques ou leurs produits. Par exemple, pour chaque CDS de la table [Genomic_Object], nous utilisons un ensemble de méthodes pour caractériser ces CDS et identifier la fonction des séquences traduites (polypeptides).

Caractérisation des CDS

Une CDS peut aussi être caractérisée par de nombreux autres attributs. Connaissant les positions de l'origine et du terminus de la réplication, nous calculons l'orientation directe ou inverse de la CDS par rapport à ces positions (table [Leading_Lagging]). La table [CodonW_Cluster] est conçue pour recevoir les résultats d'analyse statistique des CDS. Les champs CC_status et CC_coeff caractérisent la classe d'usage des codons synonymes obtenue par les analyses multivariées AFC -

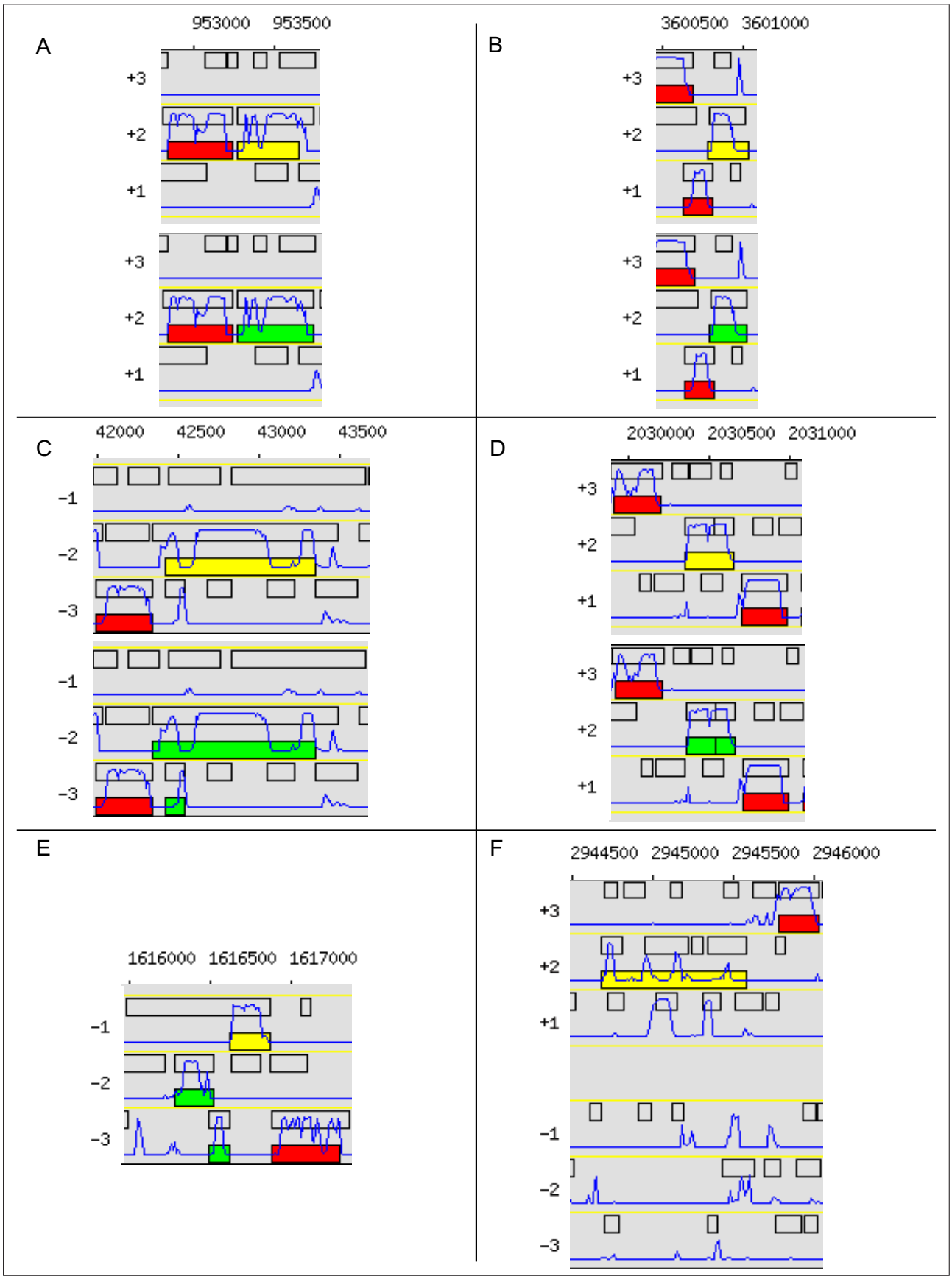


FIG. 6.2 – Différents cas de figure nécessitant la correction manuelle des bornes des CDS

FIG. 6.2 – Différents cas de figure nécessitant la correction manuelle des bornes des CDS

Les lignes jaunes représentent les six phases de lecture d'une séquence nucléique. Les rectangles transparents en position haute représentent les CDS issues du programme *prokov_orf* (l>60pb). Les courbes bleues sont les prédictions de codage issues du programme *prokov_curve*. Les rectangles rouges représentent les CDS dont le statut est à '*validated*'. Les rectangles verts représentent les CDS dont le statut est à '*checked*'. Les rectangles jaunes représentent les CDS dont le statut n'est ni à '*checked*' ni à '*validated*'.

A La carte du haut montre la CDS de *M. tuberculosis* H37Rv en jaune car elle a le statut '*no3multiple*'. La carte du bas montre la même région après correction de la position de fin de la CDS. Dans les annotations originales, ce gène est annoté comme partiel (Rv0857 953255..>953631).

B La carte du haut montre la CDS de *M. tuberculosis* H37Rv en jaune car elle a le statut '*stopInFrame*'. La carte du bas montre la même région après correction de la position de fin de la CDS. Dans les annotations originales, ce gène est annoté avec un join (Rv3224a join(3600631..3600783,3600785..3601015)).

C La carte du haut montre la CDS de *M. tuberculosis* H37Rv en jaune car elle a le statut '*noStop*'. La carte du bas montre la même région après correction de la position de début de la CDS (en vert la plus en 5′) et création d'une nouvelle CDS (en vert la plus en 3′). Dans les annotations originales, ce gène est annoté comme partiel (Rv0857 complement(<42431..43363)). La réalité de la nouvelle CDS est confirmée par Blast2p, les deux CDS s'alignent avec deux fragments différents de PR28_MYCTU (des décalages du cadre de lecture ont même été annotés pour *M. tuberculosis* H37Rv dans l'entrée Swiss-Prot).

D La carte du haut montre la CDS de *M. tuberculosis* H37Rv en jaune car elle a le statut '*multipleStop*'. La carte du bas montre la même région après correction de la position de fin de la CDS (en vert la plus en 5′) et création d'une nouvelle CDS (en vert la plus en 3′). Les annotations originales de ce gène contiennent une exception de traduction (Rv1792 /transl_except=(pos :2030519..2030521,aa :OTHER)).

E La carte montre la CDS de *M. tuberculosis* CDC1551 en jaune car elle a le statut '*checked_noStop*'. Dans les annotations originales, la longueur du gène MT1483 n'est pas multiple de trois (complement(1616248..1616836) ; '*no3multiple*'). De plus seul l'objet gene est annoté (pas d'objet CDS) et le /note= indique la présence d'un décalage du cadre de lecture authentique. La CDS en vert la plus en 3′ a été raccourcie par sa position de fin et deux nouvelles CDS ont été créées (en vert la plus en 5′ et en jaune). Dans ce cas rare et extrême, nous préférons ne pas allonger la CDS en jaune jusqu'au codon de terminaison indiqué par *prokov_orf* car cela lui ferait perdre son sens biologique (plus de 50% de la CDS serait alors artefactuelle). Les CDS qui ont le statut de cohérence '*checked_noStop*' ne doivent pas rentrer dans le processus de comparaison des annotations banque–*AMIGene* pour l'attribution d'un statut de réannotation (puisque ce processus est fondé sur la comparaison de la position du codon de terminaison d'une CDS annotée dans les banques avec celle d'une CDS prédite par *AMIGene*).

F La carte montre la CDS de *Y. pseudotuberculosis* en jaune car elle a le statut '*checked_multipleStop*'. Dans les annotations originales, la CDS or1732 est annotée comme partielle (2944676..2945572 ; /product="glucans biosynthesis protein (partial)"). De plus, on remarque la présence de quatre codons de terminaison dans le /translation=.

Ces exemples ne sont pas exhaustifs, on peut évidemment rencontrer des combinaisons de tous ces cas de figure avec plus de deux fCDS. D'une manière générale, seules les CDS qui se terminent par un codon de terminaison et qui ont une longueur supérieure à 60 pb codon de terminaison inclus, entreront dans le processus d'attribution d'un statut de réannotation.

Ether. Les autres champs (CC_) contiennent les fréquences relatives en nucléotides aux trois positions des codons calculées par le programme CodonW. De nombreuses autres fréquences relatives peuvent être déduites à partir de ces champs comme la fréquence relative en mononucléotide indépendamment de la position dans le codon ($F_R(A) = (F_R(A_1) + F_R(A_2) + F_R(A_3))/3$) ou le GC_3 ($F_R(GC_3) = F_R(G_3) + F_R(C_3)$).

Recherche de similitudes dans les banques de séquences protéiques

Le programme Blast2p de la plate-forme de comparaison de séquences, Biofacet, est utilisé pour rechercher des similitudes entre les produits de traduction des CDS de la table [Genomic_Object] et les protéines des banques publiques (*e.g.* Swiss-Prot, SWALL ; Tab. 5.2 A et B p. 172). Une description de ces banques est enregistrée dans la table [Public_Bank]. La table [Public_Protein] contient les informations sur les protéines correspondantes : le numéro d'accession (PP_accession), une description de la fonction biologique et le nom de l'organisme d'appartenance.

La table [GO_PP_CPD] décrit les résultats de recherche de similitudes. Cette table réalise une correspondance $n \leftrightarrow m$ entre les CDS (GO_id) et les protéines des banques (PP_id) : une CDS peut être similaire à plusieurs protéines des banques et inversement. Plusieurs champs permettent de caractériser une similitude comme les positions exactes de l'alignement entre les deux séquences. Les champs GOPP_score, GOPP_eval, GOPP_ident et GOPP_pos indiquent respectivement le score, la E-value, le pourcentage d'identité et de similarité entre les deux séquences. Le champ GOPP_maxlrap, rapport de la longueur de l'alignement (GOPP_matchlength) sur la plus longue des deux protéines, permet de distinguer les alignements partiels des alignements complets. Pour chaque séquence requête (produit d'une CDS de [Genomic_Object]), les dix meilleurs résultats (ordonnés selon leur E-value) sont stockés dans [GO_PP_CPD].

Reconnaissance de motifs protéiques

Une autre façon de caractériser les polypeptides est de rechercher des motifs protéiques, en particulier pour la caractérisation des domaines des protéines modulaires. Ces motifs sont recherchés avec le programme InterProScan qui utilise la base InterPro (table [InterProScan]). Par ailleurs, nous détectons les peptides signaux et les hélices alpha transmembranaires dans les séquences protéiques respectivement à l'aide des programmes SignalP et TMHMM (resp. [SignalP] et [TMHMM]).

Classes fonctionnelles de génomes modèles

Pour attribuer des classes fonctionnelles aux polypeptides, nous utilisons le système de classification multi-fonctionnelle d'*E. coli* K-12 fourni par la banque *GenProtEC* [Serres *et al.*, 2004] et intégré dans la table [MultiFun_Classif]. *La table de correspondances [GO_MC_CPD] permet d'attribuer plusieurs classes fonctionelles à un même polypeptide.* Par exemple, le module b0149_1 est impliqué dans la biosynthèse du peptidoglycane (métabolisme et structure cellulaire), dans la division cellulaire (processus cellulaire) et situé dans la membrane interne (localisation). Le module

b0149_2 est impliqué dans la biosynthèse du peptidoglycan, dans les processus cellulaires de la division cellulaire et de la résistance aux antibiotiques (*e.g.* penicilline), et localisé dans le cytoplasme.

Dans le cadre de l'analyse de l'usage des codons synonymes (voir p. 197), les partitions en k classes que nous obtenons sont souvent corroborées par des propriétés biologiques particulières : biais mutationnel (une classe de gènes préférentiellement sur le brin précoce contre une classe de gène préférentiellement sur le brin tardif), l'essentialité, l'expressivité, l'origine exogène, etc. Nous collectons donc des données de référence sur ces classes fonctionnelles, au moins dans le cas d'organismes modèles comme *E. coli* K-12 et *B. subtilis* (gènes exogènes [GenProtEC], gènes essentiels [Predicted_essential] et gènes hautement exprimés [Highly_eXpressed] ; [Wei *et al.*, 2001, Karlin *et al.*, 2001] et Tab. 5.2 A p. 172).

Familles fonctionnelles des génomes complets

La table [COG] regroupe les 3307 classes de la banque de COG identifiées par COG_id et dont la fonction est décrite dans le champ COG_function. Une CDS pouvant être classée dans un ou plusieurs COG, la table [CO_COG_CPD] établit des correspondances $n \leftrightarrow m$ entre les tables [Genomic_Object] et [COG]. Les correspondances sont calculées par le programme COGnitor : soit nous récupérons les données de la banque de COG qui comprend 43 génomes, soit nous exécutons une version locale du COGnitor sur un protéome (Tab. 5.2 p. 172). La table [COG_Gene] permet d'établir des correspondances entre les COG_id (*e.g.* COG0007) et les CG_gene_name (*e.g. cysG*) non spécifique de chaque organisme. La table [COG_Bank] permet d'établir des correspondances entre les COG_id et les CB_label (*e.g.* Rv0511) spécifiques de chaque organisme (les CB_COG_label comportent en plus la notion de module, *e.g.* Rv0511_1).

Prédiction de fonctions enzymatiques

Les données de la banque ENZYME sont chargées dans la table [Enzyme] [Bairoch, 2000]. Les réactions enzymatiques sont identifiées par le champ EC_id qui correspond à un numéro attribué par l'*Enzyme Commission* (EC). La méthode PRIAM (PRofils pour l'Identification Automatique du Métabolisme [Claudel-Renard *et al.*, 2003, Claudel-Renard, 2003]) permet d'effectuer une recherche de similitude de chaque séquence protéique requête d'un protéome complet contre une banque de profils enzymatiques, en utilisant le programme RPS-BLAST (*Reverse Position-Specific Basic Local Alignement Search Tool* [Altschul *et al.*, 1997]). La banque est construite préalablement : à chaque numéro EC correspond une ou plusieurs matrices de score position spécifique (PSSM). Ces PSSM sont calculées à partir des alignements multiples de toutes les séquences de Swiss-Prot qui possèdent le même numéro EC (annotations de la banque Enzyme) par le programme PSI-BLAST (*Position-Specific Iterated BLAST* [Marchler-Bauer *et al.*, 2003]). PRIAM va donc attribuer aux produits des CDS stockées dans [Genomic_Object], zéro (s'il ne s'agit pas d'un enzyme), une (s'il s'agit d'une monoenzyme) ou plusieurs (s'il s'agit d'un multienzyme) activités enzymatiques. Ces résultats sont stockés dans la table [GO_EC_CPD] (*e.g.* le numéro du profil et le numéro EC sont contenus

respectivement dans les champs GOEC_profil et EC_id).

Recherche d'orthologues, de paralogues

Nous avons déjà défini les relations d'orthologie et de paralogie entre deux gènes dans la partie de l'état de l'art (voir p. 51). Dans le but de comparer les génomes deux à deux, on définit des relations de *correspondance* entre les objets génomiques d'un génome G_1 et ceux d'un génome G_2 (1 et 2 pouvant être deux espèces procaryotes ou deux souches d'une même espèce). Ces relations de *correspondance* peuvent être fondées sur divers critères de comparaison (*e.g.* recherche de similitudes contre une banque de séquences protéiques ou contre une banque de motifs protéiques).

La table [GO_GO_CPD] permet de conserver les résultats de comparaisons de séquences entre les CDS de deux génomes pour la recherche d'orthologues, ou entre les CDS d'un même génome pour la recherche de paralogues. Elle a la même structure que la table [GO_PP_CPD] à ceci près qu'elle établit une relation de similitude entre deux objets génomiques de type CDS (GO_id_1 et GO_id_2). Deux types de *correspondances* sont particulièrement intéressantes : les *BBH* (*Bidirectional Best Hit*, *i.e.* les meilleurs alignements bidirectionnels) et les *BH* (*Best Hit*, *i.e.* les meilleurs alignements) [Overbeek *et al.*, 1999]. Deux gènes X_1 et X_2 de deux génomes G_1 et G_2 sont en *BBH* si X_1 et X_2 ont une similitude de séquence, s'il n'existe pas un gène Y_2 de G_2 plus similaire à X_1 que X_2 et s'il n'existe pas un gène Y_1 de G_1 plus similaire à X_2 que X_1. Deux gènes X_1 et X_2 sont uniquement en *BH* si l'une des deux dernières conditions n'est pas respectée. Ces deux notions permettent de supposer que deux gènes en *BBH* sont plus vraisemblablement des orthologues que deux gènes en *BH*. La table [GO_GO_CPD] comporte les champs GOGO_order_1 et GOGO_order_2 qui permettent d'ordonner les alignements suivant un critère (*e.g.* E-value, score). Si les champs GOGO_order_1 pour GO_id_1 *et* GOGO_order_2 pour GO_id_2 sont égaux à 1, alors GO_id_1 et GO_id_2 sont en *BBH*. Si uniquement GOGO_order_1 *ou* GOGO_order_2 est égal à 1, alors GO_id_1 et GO_id_2 sont en *BH*.

La table [GO_GO_CPD] permet donc de lier des orthologues putatifs d'un génome G_1 à un génome G_2. L'absence de relation de X_1 vers G_2 au sein de [GO_GO_CPD] peut suggérer la spécificité de ce gène (X_1 est unique à G_1 par rapport à G_2) sans toutefois l'affirmer. La table [GO_GO_CPD] contient également le champ SY_id (identifiant unique de groupe de synténie) qui permet de renseigner l'implication d'une relation de *correspondance* entre deux gènes au sein d'un groupe de synténie (voir p. 192).

6.3.3 Prédictions relationnelles

Les tables d'annotation relationnelle permettent d'établir des relations complexes entre les objets génomiques ou entre leurs produits afin de reconstituer des unités biologiques (*e.g.* unités de transcription, complexes protéiques, voies métaboliques, cascades de signalisation ; Fig. 1.1 p. 35).

Prédictions d'ilôts génomiques

Il existe plusieurs termes pour désigner les gènes acquis par transfert horizontal : on parle de gènes HGT (*Horizontal Gene Transfer* [Lawrence & Hendrickson, 2003]), de gènes LGT (*Lateral Gene Transfer* [Daubin *et al.*, 2003b]), de gènes pA (*putative Alien* [Karlin, 2001]), de gènes xénologues [Fitch, 2000], de gènes d'origine étrangère, exogène ou extra-chromosomique [Serres *et al.*, 2004]. Un îlot génomique (*Genomic Island* (GI)) est un groupe de gènes colocalisés sur le chromosome, acquis par transfert horizontal, essentiels pour la plasticité adaptative (*fitness*) et la survie des bactéries à des conditions de stress (*e.g.* opéron cobalamine de *S. enterica* serovar Typhimurium LT2, système de capture du fer des *Yersinia* spp. [Hacker & Carniel, 2001]). Les îlots de pathogénie (*PAthogenicity Island* (PAI)) et de surcroît les îlots de haute pathogénie (*High Pathogenicity Island* (HPI)) sont des îlots génomiques dont la fonction est impliquée dans la virulence des bactéries pathogènes (*e.g.* adhésine, invasine, système de secrétion de type III ou IV, toxines [Karlin, 2001]).

Ainsi, une méthode pertinente pour la prédiction d'îlots génomiques consiste à rechercher des groupes de gènes HGT colocalisés sur le chromosome. Toute la difficulté réside dans la prédiction des gènes HGT. En effet, il existe un débat sur les gènes HGT : certains ont observé que les gènes atypiques dans leur usage des codons synonymes (*i.e.* classe III AT_3 riche chez *E. coli* K-12 et *B. subtilis*) étaient souvent groupés sur le chromosome et présentaient des similitudes avec des gènes connus pour être transférés horizontalement (*e.g.* gènes de phages, de toxines [Moszer *et al.*, 1999]). Cependant, d'autres ont annotés des îlots génomiques GC_3 riches (Fig. 11.2 p. 343). Enfin, d'autres encore pensent que les gènes transférés horizontalement n'ont pas forcément un usage des codons atypique et inversement [Koski *et al.*, 2001]. L. Koski *et coll.* invoquent plusieurs raisons.

- Si l'événement est trop ancien, l'usage des codons synonymes des gènes acquis a eu le temps de s'adapter à celui de gènes typiques de la bactérie.
- Si l'événement a eu lieu entre deux espèces bactériennes ayant des usages des codons synonymes similaires, alors il sera impossible d'observer une différence significative entre l'usage des codons synonymes des gènes acquis et celui des gènes typiques.
- D'autres pressions de sélection peuvent expliquer la présence de gènes avec un usage des codons synonymes atypique.

Comme nous le verrons p. 226, la composition des gènes HGT serait plus influencée par des propriétés structurales de l'ADN que par la composition du génome donneur. La classe III permettrait donc de repérer une partie des transferts horizontaux : les transferts *récents* de gènes AT_3 riches. En attendant d'accueillir des résultats d'une nouvelle méthode de prédiction d'îlots génomiques (voir p. 353) et/ou des données d'autres ressources (*e.g.* HGT-DB [Garcia-Vallve *et al.*, 2003] IslandPath [Hsiao *et al.*, 2003] ; voir p. 338), la table [Genomic_Island] de PkGDB contient des GI de référence comme les PAI de *Y. pestis* CO92 [Parkhill *et al.*, 2001b] et de *P. luminescens* [Duchaud *et al.*, 2003].

Reconstruction de voies métaboliques

Les numéros EC attribués par PRIAM aux CDS présentant une activité enzymatique sont le point de départ de la reconstruction de voies métaboliques. Une première étape permet de visualiser sur les graphes métaboliques d'un organisme modèle le contenu enzymatique d'un organisme en cours d'étude. La démarche consiste à réaliser des interconnexions entre PkGDB et deux bases de données métaboliques : KEGG et BioCyc. Le serveur KEGG rassemble toutes les voies métaboliques possibles des génomes complets. Il est capable de représenter des voies métaboliques à partir de numéros EC. La base BioCyc (MetaCyc et EcoCyc) permet de faire des requêtes sur les voies métaboliques d'un nombre plus limité d'organismes.

Dans une seconde étape, PkGDB permet, à partir des numéros EC *et* des orthologues *putatifs* entre l'organisme modèle et l'organisme étudié, d'obtenir les listes d'enzymes communes aux deux organismes, uniques à l'organisme modèle et à l'organisme étudié. Ces listes sont calculées dynamiquement selon des paramètres de similitude de séquences, et vont permettre de redessiner les schémas métaboliques de l'organisme modèle en distinguant les réactions enzymatiques présentes ou absentes chez l'organisme étudié. La base de données KEGG possède un outil permettant de colorer, sur les graphes métaboliques, les numéros EC suivant un code couleur spécifié : il peut donc servir à visualiser les réactions communes et uniques et ainsi à localiser des points de ruptures dans les voies métaboliques. Ces ruptures indiquent (1) que les enzymes *manquantes* sont bien présentes dans l'organisme étudié mais qu'elles n'ont pas été détectées par les méthodes de prédiction fonctionnelle, (2) que l'organisme étudié n'est pas capable de synthétiser les métabolites en aval d'une rupture, ou (3) que l'organisme étudié utilise une voie métabolique alternative pour la synthèse de ces métabolites.

Prédictions de groupes de synténie

Dans le but d'aider à l'annotation fonctionnelle des gènes, nous nous intéressons à la détection de groupes de gènes dont l'organisation reste relativement conservée entre deux génomes procaryotes : on parle de groupes de synténie. En effet, l'observation de telles structures peut traduire un éventuel couplage fonctionnel entre les produits des gènes concernés, comme c'est le cas entre les produits de gènes appartenant à des opérons ou à des régulons procaryotes. Un groupe de synténie est un ensemble de couples d'orthologues putatifs dont les positions relatives restent *voisines* sur les génomes G_1 et G_2 ; autrement dit, l'organisation locale des gènes est conservée.

Le *Syntonizer* détecte les groupes de synténie en comparant les génomes par paires, à la recherche des groupes de gènes *colocalisés* sur un génome dont les *correspondants* restent *colocalisés* sur l'autre génome. La méthode du *Syntonizer* modélise les groupes de synténie en utilisant le formalisme des graphes mathématiques. Dans cette représentation, les gènes sont les nœuds du graphe et sont connectés par deux types d'arêtes : les relations de *correspondance* (inter-génome) et les relations de *colocalisation* (intra-génome). Les relations de *correspondance* sont principalement établies sur des résultats de comparaison de séquences avec des contraintes restrictives sur les résultats de

similitude (*e.g. BBH*, *BH*, pourcentage d'identité, rapport des longueurs de l'alignement sur la plus courte des deux séquences). Ces relations peuvent aussi représenter l'appartenance à des classes fonctionnelles proches (COG, numéro EC) ou encore traduire la présence de domaines protéiques conservés (InterPro). Une des particularités de cette méthode est qu'elle gère les *correspondances multiples* entre les gènes autorisant ainsi la détection de fusions–fissions, et autres duplications de gènes (on ne se limite donc pas à la définition du *BBH* en terme de comparaison de séquence). Les relations de *colocalisation* sont définies par un paramètre de *gap* (nombre maximum de gènes consécutifs non membres du groupe, séparant deux membres du groupe). Cette définition autorise tous les remaniements possibles (inversion, translocation, insertion–délétion) au sein d'un groupe de synténie. L'algorithme implémenté dans le *Syntonizer* calcule les *groupes maximaux d'arêtes de correspondance* dont les extrémités, sur chacun des deux génomes comparés, forment une sous-composante connexe selon la relation de *colocalisation*. Les groupes sont obtenus par un raffinement de partitions sur l'ensemble des arêtes de *correspondance* présentes dans le graphe. L'originalité de cette approche réside dans le fait que l'utilisateur est libre de définir les deux types de relations et de calculer dynamiquement les groupes de synténie qui en résultent.

Les groupes de synténie prédits par le *Syntonizer* sont stockés dans la table [GO_Synteny] de la base PkGDB. Cette table donne pour chaque groupe de synténie : le couple de génomes concernés (champs S_id_1 et S_id_2), les positions de début et de fin du groupe de synténie sur chacun des deux génomes (SY_begin_1, SY_end_1, SY_begin_2 et SY_end_2), le nombre de gènes impliqués (SY_gene_nb_1 et SY_gene_nb_2 pouvant différer dans le cas de *correspondance multiple*). Un groupe de synténie est un ensemble d'arêtes de *correspondance* contenues dans la table [GO_GO_CPD] et une arête de *correspondance* ne peut être impliquée que dans un seul groupe de synténie (correspondance $n \leftrightarrow 1$ entre [GO_GO_CPD] et [GO_Synteny]). Le lien entre un groupe de synténie de [GO_Synteny] et les gènes de [Genomic_Object] qui en font partie se fait donc par l'intermédiaire de la table [GO_GO_CPD] qui associe les identifiants des deux orthologues (*i.e.* GO_id_1 et GO_id_2) à l'identifiant du groupe de synténie (*i.e.* SY_id).

Lors de l'annotation fonctionnelle d'un gène, la connaissance des groupes de synténie permet de replacer le gène dans son contexte génomique au lieu de le considérer individuellement.

Ces annotations relationnelles, voies métaboliques, ilôts génomiques, et groupes de synténie sont accessibles dans l'interface cartographique *MaGe* adossée à PkGDB.

6.4 Réconciliation des annotations des banques et des prédictions de l'A.G.C.

Pour un génome, il s'agit de :

1. comparer les annotations des banques et les prédictions de l' A.G.C. (Atelier de Génomique Comparative) pour déterminer les objets communs aux deux types d'annotation, uniques aux annotations des banques et uniques aux prédictions de l'Atelier de Génomique Comparative,

2. attribuer un statut de réannotation à certaines des CDS uniques,
3. remplir la table centrale de PkGDB [Genomic_Object] avec les objets génomiques reconciliés issus des annotations des banques et/ou des prédictions de l'Atelier de Génomique Comparative.

Il est possible d'étendre ce concept de réconciliation des annotations à d'autres types d'objet que les CDS, comme les ARNt ou les ARNr. La gestion de cette partie de la base PkGDB, réconciliation des deux jeux de CDS et chargement de [Genomic_object], est complètement automatisée à travers un script et nécessite d'avoir préalablement construit les matrices de transition à partir de l'étude de l'usage des codons synonymes des CDS des banques dont les bornes ont été soigneusement vérifiées. Ce script enchaîne différentes étapes pour un génome :

1. Vérification de la complétion des corrections manuelles des CDS des banques et de la cohérence des corrections (voir p. 183).
2. Calcul de la probabilité moyenne de codage des CDS des banques contenues dans [Compare_Annotation] (voir le programme *GBK_max_Pc* p. 276) et mise à jour des champs CA_Pc et CA_CU_matrix.
3. Prédiction des CDS *AMIGene* et chargement de la table [AMIGene] (voir p. 237).
4. Sur la base de la comparaison de la position du codon de terminaison des CDS des banques de [Compare_Annotation] et des CDS d'[AMIGene], mise à jour du champ CA_status ('*common*' ou '*uniqBank*') des CDS des banques et chargement des CDS uniques à *AMIGene* dans [Compare_Annotation] (CA_status = '*uniqAGC*' ; voir p. 275).
5. Extraction des listes de CDS uniques aux annotations des banques et aux prédictions *AMIGene* et traduction des séquences (voir p. 275).
6. La plate-forme Biofacet permet d'effectuer une recherche de similitude des jeux de polypeptides uniques contre la banque des séquences protéiques SWALL (voir p. 275).
7. Attribution d'un statut de réannotation à certains des gènes uniques en fonction de plusieurs critères (*e.g.* longueur, probabilité moyenne de codage, recouvrement, similitude) et mise à jour de CA_status ('*suspiciousBank*', '*wrongBank*', '*ambiguousAGC*' et '*newAGC*' ; voir p. 275).
8. Alimentation de la table [Genomic_object] avec les CDS et fCDS « curées », les ARNt et les ARNr de la table [Compare_Annotation] et d'autres objets génomiques de la table [Databank_Feature]. Lors du remplissage de [Genomic_Object], quatre cas sont traités :
   – Les objets génomiques des banques de [Databank_Feature] dont le DF_id est orphelin (*i.e.* il n'a pas de DA_id correspondant dans la table [DA_DF_CPD]). Ils sont par exemple de type : RBS, misc_RNA, sc_RNA, misc_feature, repeat_unit, etc. La table de correspondance [GO_DF_CPD] permet d'accéder directement aux annotations originales, en particulier aux attributs et aux objets qui n'ont pas été propagés dans les tables [Databank_Annotation] et [Compare_Annotation].

- Les fCDS des banques de [Compare_Annotation]. La table [GOCDS_GO_CPD] permet de construire des CDS de type complexe (cCDS) constituées de plusieurs fragments décalés dans leur phase de lecture (fCDS).
- Les autres objets génomiques des banques de [Compare_Annotation] : les CDS communes et uniques, les ARNt et les ARNr.
- Les CDS de [Compare_Annotation] uniques à *AMIGene* (CA_status = *'uniqAGC'*).

Ainsi, si une CDS contient une mutation (*e.g.* un décalage du cadre de lecture) son champ GO_mutation ne peut être à 'no'. Si le gène muté est composé d'une CDS, alors le GO_type de l'objet CDS contiendra la valeur *CDS*. En pratique, ce cas peut correspondre à un fragment de CDS isolé (les autres fragments ont-ils été supprimés par une délétion ou se trouvent-ils plus loin sur le chromosome ?). Si le gène muté est composé de plusieurs fragments de CDS colocalisés, alors le GO_type de ces objets seront à 'fCDS' (et toute fCDS est liée à une cCDS).

6.5 Instances de PkGDB

Partant de la structure générique du modèle relationnel PkGDB que nous venons de décrire, différentes instances ont été crées. Lors de l'instanciation du modèle, les langages de programmation utilisés sont MySQL pour communiquer avec la base, des langages Unix (Bash et Shell) pour enchaîner des lignes de commande et manipuler des fichiers, le Awk et le Perl pour « parser » et formater les données des fichiers et enfin le C pour les calculs. Généralement, les programmes ne modifient pas la base directement, mais génèrent plutôt un fichier que l'utilisateur chargera dans la base après vérification. Ce fichier comporte un enregistrement par ligne où chaque champ est séparé par le caractère '$'. L'intégration de ces données dans PkGDB est alors réalisée avec la commande appropriée[7]. Cette procédure n'est pas réalisée dans le cas des tables de correspondance et des tables d'entités associées, car l'ensemble nécessite un chargement en deux temps. Par exemple, à chaque fois que l'on veut insérer une nouvelle entrée dans [Databank_Annotation], il faut en même temps la relier aux différents DF_id dont elle est issue. Autrement dit, un programme insère d'abord une nouvelle entrée dans [Databank_Annotation], puis une ou plusieurs nouvelles entrées dans [DA_DF_CPD] (le LAST_INSERT_ID() précédemment créé et un DF_id qui lui correspond).

La structure logique représentée dans la figure 6.1 p. 176 est donc utilisée pour construire plusieurs bases de données : (i) une base de données multigénomes, PkGDB dont la vocation est de rassembler les données de génomes bactériens publiés et réannotés (Tab. 10.1 p. 293), et les données de génomes que nous étudions au laboratoire. Ce pool de données, qui sera mise à la disposition de la communauté scientifique dans un très proche avenir, sert de point de départ à la construction d'autres bases de données plus spécialisées : (ii) des bases de données spécialisées dont la thématique est l'annotation ou la réannotation de micro-organismes particuliers. Nous avons aujourd'hui deux bases spécialisées : *Acineto*DB et *Entero*DB. *Acineto*DB a été construite pour le

[7] echo "LOAD DATA LOCAL INFILE 'chemin du fichier' INTO TABLE 'Nom_Table' FIELDS TERMINATED BY '$'" | mysql -u agc -p ****** -h masaya8 –local-infile=1 pkgdb

projet d'annotation du génome d'*Acinetobacter* ADP1 séquencé au Genoscope [Barbe *et al.*, 2004]. Dans le cadre de ce projet, les CDS prédites par *AMIGene* de la table [Genomic_Object] sont validées manuellement en utilisant l'interface d'annotation *MaGe*. *Entero*DB sert à la réannotation et à l'exploration des génomes d'entérobactéries pathogènes comme *E. coli* O157:H7 EDL933 [Perna *et al.*, 2001], *S. enterica* serovar Typhimurium LT2 [McClelland *et al.*, 2001], *S. enterica* serovar Typhi CT18 [Parkhill *et al.*, 2001a], *Y. pestis* CO92 [Parkhill *et al.*, 2001b], *Y. pseudotuberculosis* [Chain *et al.*, 2004], *P. luminescens* [Duchaud *et al.*, 2003], en référence au génome de l'entérobactérie commensale *E. coli* K-12 [Rudd, 2000]. Trois autres bases sont en cours de construction : HaloplanktisDB, StaphyloDB et PseudomonasDB. L'objectif de ces bases consiste donc à rassembler, autour du ou des génomes d'intérêt (*i.e.* dont le processus de (ré)annotation doit être réalisé), les données des génomes que l'on souhaite comparer en utilisant d'une part les données publiques stockées dans PkGDB (donc corrigées et enrichies), et d'autre part les données de bases de référence (*GenProtEC*, RegulonDB, etc.).

Ma participation à la gestion des instances PkGDB et *Entero*DB est importante : j'ai développé et exécuté différentes procédures destinées à l'alimentation, à la correction, à la mise à jour des annotations de génomes publics entièrement séquencés. Une base telle que PkGDB est en évolution permanente (*e.g.* intégration de données issues de nouveaux outils d'analyse, de nouveaux organismes). Nous sommes en cours de développement d'une interface d'interrogation.

Chapitre 7

Apprentissage des séquence d'ADN par des chaînes de Markov : *AMIMat*

Dans le cadre d'un projet d'annotation d'un nouveau génome procaryote ou de réannotation d'un génome déjà publié, deux stratégies complémentaires sont utilisées pour prédire les CDS :

1. *AMIGene Matrices* (*AMIMat*) est une stratégie semi-automatique d'apprentissage des séquences codantes et non-codantes.
2. *Annotation of Microbial Genes* (*AMIGene*) est une stratégie automatique de prédiction de CDS (voir p. 237).

Ces deux stratégies sont fondées sur le modèle statistique des chaînes de Markov, couramment utilisé pour modéliser les séquences d'ADN (voir p. 86). L'application de ce modèle aux séquences d'ADN repose sur l'hypothèse que les contraintes de codage induisent un biais statistique dans la distribution des oligonucléotides (*e.g.* codons). En pratique, les deux stratégies reposent sur les modules du programme de prédiction de CDS par chaînes de Markov, *Prokov* (voir p. 96).

7.1 Pourquoi un autre programme de prédiction de gènes bactériens ?

La recherche de CDS dans les génomes procaryotes consiste généralement à mettre en œuvre des méthodes intrinsèques ou *ab initio*. Les méthodes intrinsèques se fondent uniquement sur les propriétés locales de la séquence à annoter : composition en oligonucléotides et signaux. Même si la prédiction de gènes chez les procaryotes est en théorie plus facile que chez les eucaryotes, des problèmes restent non résolus (choix du « vrai » codon d'initiation de la traduction ou prédiction du début de la transcription) et les méthodes imparfaites (équilibre entre faux-positifs et faux-négatifs). Nous retiendrons les principales difficultés suivantes qui ont été, en partie, traitées dans le cadre de ce travail :

1. Les génomes bactériens sont compacts car 80 à 90% de leur séquence correspond à des gènes

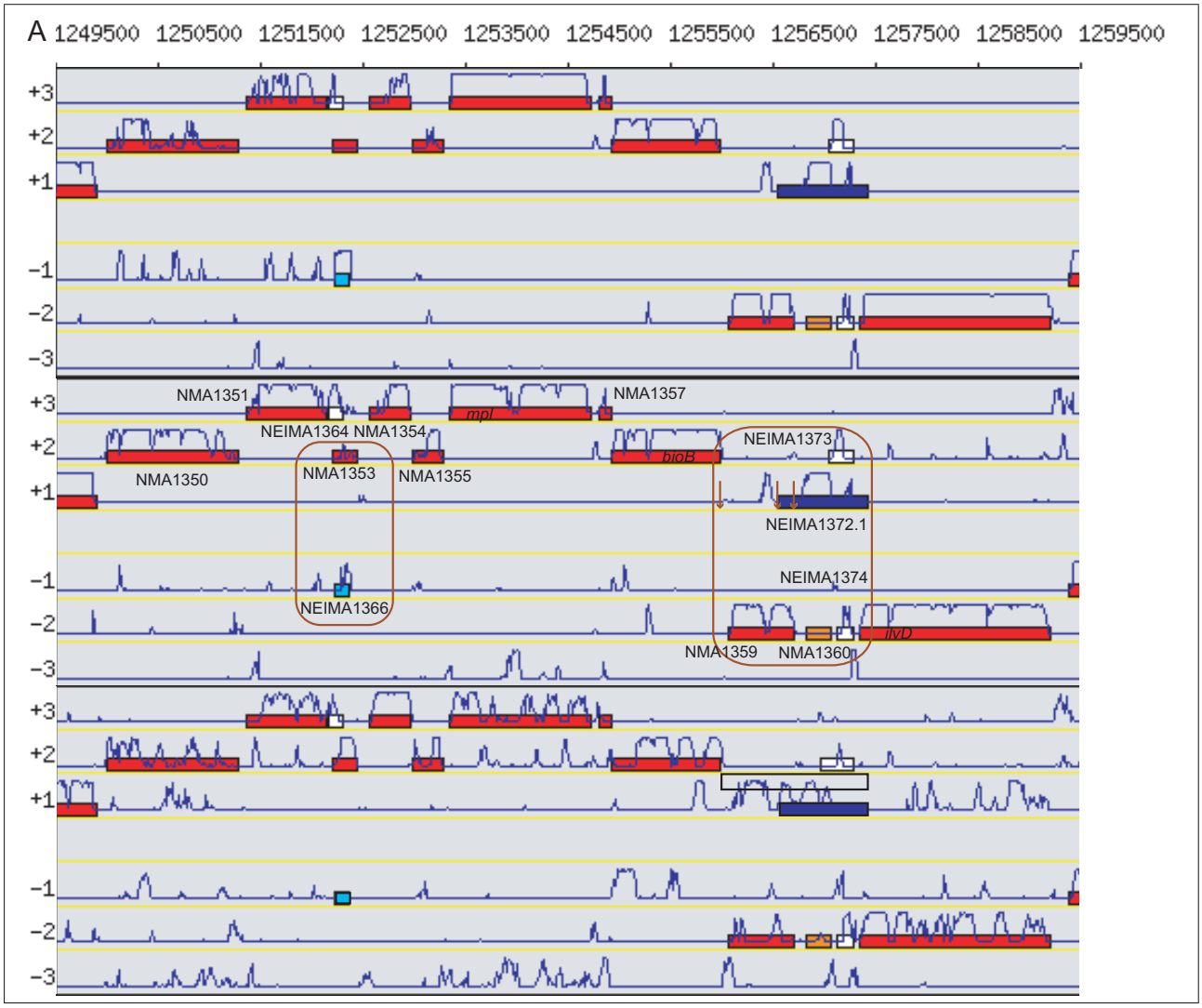


FIG. 7.1 – A) Région difficile à annoter chez *N. meningitidis* Z2491 (Serogroup A)
Les trois cartes représentent le même fragment de séquence. Les courbes de probabilité de codage de la première, de la deuxième et de la troisième carte ont été calculées respectivement avec la matrice I, la matrice II et la matrice III. Le rectangle transparent de la troisième carte représente la CDS prédite par *prokov_orf* (LS). Les CDS NMA1350 (protéine du cycle cellulaire), NMA1351 (ARN-méthyltransférase) et NEIMA1364 (fragment de NADH-déshydrogénase) répondent mieux avec la matrice II. Les CDS NMA1353 (lipoprotéine qui possède une région répétée) et NMA1354 (contenu en G+C anormalement bas) répondent mieux avec la matrice III. La CDS NEIMA1366 répond mieux avec la matrice I et possède une région répétée mais pas de similitude. Au vue de cette analyse, il semble que la CDS NMA1353 soit une CDS atypique dans son usage des codons synonymes tandis que la NEIMA1366 serait une CDS fantôme due aux répétitions. Les CDS NMA1359 (biosynthèse de la capsule), NMA1372.1 (biosynthèse de la capsule ; possède des régions répétées) et NMA1373 (biosynthèse de la capsule) répondent mieux avec la matrice I. La première, la deuxième et la troisième flèche au niveau de NEIMA1372.1 indiquent respectivement le codon d'initiation le plus en 5′, celui prédit par *AMIGene* et celui qui est le plus vraisemblable. Aucune de trois matrices ne permet de prédire correctement la CDS 1360 sans similitude. Au vu de cette analyse, les CDS NMA1360 et NEIMA1374 semblent artefactuelles tandis que NEIMA 1372.1 et NEIMA1374 semblent être les vestiges d'une duplication ancestrale (le gène fonctionnel de biosynthèse de la capsule étant NMA1359).

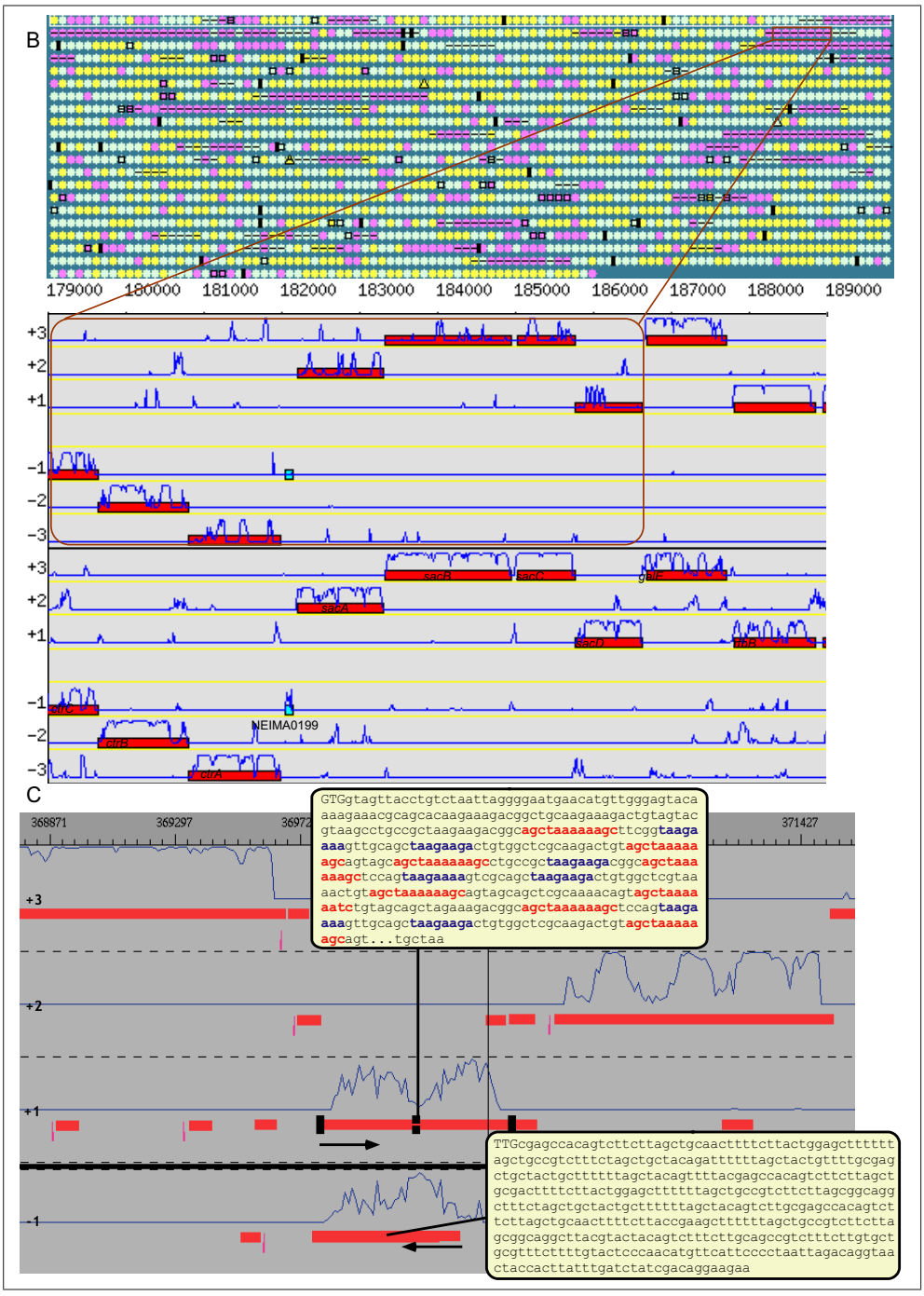


FIG. 7.1 – Région atypique et répétitions

B) Région atypique de *N. meningitidis* Z2491 (Serogroup A). La première image a été téléchargée à partir du site d'IslandPath [Hsiao *et al.*, 2003]. Pour la légende, voir p. 296. Les courbes de probabilité de codage de la première et de la seconde carte ont été calculées respectivement avec la matrice I et la matrice III. Quatre CDS (*sacA*, *sacB*, *sacC* et *sacD*) sont ratées par *AMIGene* lorsqu'on utilise qu'une seule matrice de transition. Cependant, elles n'avaient ni le statut '*suspiciousBank*' ni le statut '*wrongBank*', car leur produit est similaire à des protéines de biosynthèse de la capsule.

C) CDS fantômes et répétitions chez *Chlamydia trachomatis*. Ici, la CDS en phase +1 est celle qui est annotée par les auteurs, mais *AMIGene* a sélectionné celle de la phase -1 située à la même position du génome (bien qu'elle soit plus courte, cette CDS a une probabilité moyenne de codage plus élevée). En regardant la séquence nucléique de la CDS en phase +1, il est apparu qu'elle était constituée d'au moins 7 répétitions de l'oligonucléotide AGCTAAAAAGC (motif rouge), et de plusieurs répétitions TAAGAAGA moins conservées (motif bleu). Sur le brin complémentaire, ces répétitions sont naturellement retrouvées et dans le cas présenté ici, une phase ouverte suffisamment longue a été détectée par *AMIGene*, ce qui conduit au cas d'une inclusion en sens contraire, entre deux CDS de Pc significative.

(d'ARN et de protéines). De plus, la structure en opéron des génomes bactériens autorise de courts chevauchements entre CDS (une dizaine de codons au maximum), ce qui n'est pas toujours pris en compte dans les méthodes de prédiction de gènes (voir p. 114). Il existe bien évidemment des exceptions, comme le petit gène *comS* inclus dans la CDS *sfrAB*, de même orientation (directe) et codant un polypeptide de 46 aa nécessaire au développement de la compétence chez *B. subtilis* [Hamoen *et al.*, 1995].

2. A cause de la compacité de ces génomes, il est difficile de définir un jeu suffisamment conséquent de séquences non-codantes natives.

3. Définir les positions des CDS implique de trouver un codon d'initiation et un triplet de terminaison de la traduction. Il n'y a pas d'ambiguïté possible dans le choix du triplet de terminaison, puisque ces triplets ne codent pas d'acide aminé. A l'inverse, le choix du codon d'initiation est problématique puisque les codons ATG, GTG et TTG servent à la fois à initier la traduction et à coder les acides aminés méthionine, valine et leucine respectivement. Un modèle simple consiste à choisir le codon d'initiation qui génère la CDS la plus longue (*Leftmost Start* (LS) ou codon d'initiation le plus en 5′). Ce modèle ne décrit pas fidèlement la réalité car, statistiquement, on s'attend à devoir réajuster le codon d'initiation dans approximativement 25% des cas. Plus précisément, si l'on regarde en 5′ du « vrai » ATG, on a une chance sur quatre de tomber d'abord sur un ATG et trois chances sur quatre de tomber d'abord sur un TAA, un TAG ou un TGA [Lukashin & Borodovsky, 1998]. Si l'on effectue le même raisonnement en prenant en compte les trois codons d'initiation ATG, TTG et GTG, alors on a une chance sur deux de tomber d'abord sur un codon d'initiation et une chance sur deux de tomber d'abord sur un codon de terminaison. Effectivement, plus de 40% des codons d'initiation ont été réajustés dans le cas des dernières réannotations d'*E. coli* K-12, *B. subtilis* et *M. tuberculosis* H37Rv (Tab. 7.1 p. 213). Si le LS ne correspond pas au « vrai » codon d'initiation, la CDS générée sera trop longue, ce qui peut engendrer un recouvrement (*e.g.* un chevauchement, voire une inclusion) avec une ou plusieurs CDS adjacentes, quel que soit leur sens de transcription (Fig. 7.1 A p. 198).

4. Les décalages du cadre de lecture des CDS doivent également être pris en compte. L'origine de ces sauts de phase est soit artefactuelle (erreur de séquençage), soit naturelle (*frameshift* authentique). Dans le cas d'une CDS constituée par exemple de deux fragments, le fragment en 5′ doit, en théorie, contenir un codon d'initiation, mais pas forcément un codon de terminaison et *vice versa* pour le fragment en 3′. En pratique, chaque CDS (même si ce n'est qu'un fragment) doit être définie par un codon d'initiation et de terminaison ce qui peut générer un recouvrement important entre les deux fragments, allant parfois jusqu'à l'inclusion (Fig. 10.8 p. 325). Ce phénomène est amplifié par le choix du codon d'initiation le plus en 5′.

5. Selon les génomes, les études portant sur l'hétérogénéité des CDS ont permis de mettre en évidence deux, trois ou quatre types de composition (Fig. 7.1 B p. 198 et voir p. 212). Les CDS riches en A+T, dites atypiques, sont particulièrement difficiles à prédire car cette carac-

téristique compositionnelle les rapproche de celle de l'intergénique [Hayes & Borodovsky, 1998b, Nicolas *et al.*, 2002]. Les méthodes de prédiction de gènes doivent donc prendre en compte les différents types de composition des CDS.

6. Les génomes riches en G+C comme *M. tuberculosis* H37Rv sont plus difficiles à annoter que les génomes pauvres en G+C mais pour d'autres raisons. En effet, ces génomes ont moins de codons de terminaison, ce qui génère plus de longues ORF artefactuelles recouvrant de « vraies » CDS (Tab. 8.1 A p. 253 et Fig. 8.1 B p. 240). Les cas de réajustement du codon d'initiation des CDS sont alors plus fréquents (Tab. 7.1 p. 213). La partie $5'$ non-codante d'une CDS ou d'un fragment, dont on a choisi le codon d'initiation le plus en $5'$, est en moyenne plus longue. Ces caractéristiques accentuent aussi la taille et le nombre de recouvrements liés au choix du codon d'initiation ou aux décalages du cadre de lecture.

7. La présence de répétitions, comme celles observées dans les génomes des *Neisseria* ou des *Mycoplasmes*, augmente la difficulté d'annotation. En effet, il arrive que des répétitions dans une CDS expliquent la présence d'une ombre du codant sur le brin opposé (Fig. 7.1 C p. 198 et voir p. 93), que des répétitions en début de gène produisent des codons d'initiation alternatifs, que des IS interrompent des CDS, que des duplications génèrent des paralogues ou soient à l'origine d'autres réarrangements, etc.

8. La détection de très petits gènes qui ne contiennent pas suffisamment d'information pour que leur reconnaissance par chaînes de Markov soit statistiquement significative. Le biais de codage d'une CDS dont la longueur est inférieure à 63 pb (codon de terminaison inclus) peut être alors masqué par un biais d'échantillonage quelconque [Nicolas, 2003]. La plus petite protéine enregistrée dans Swiss-Prot qui ne soit pas un fragment est la lichenine (bacteriocine) de *Bacillus licheniformis* (P82907 a une longueur de 12 aa).

Finalement, les régions génomiques contenant des gènes impliqués dans la virulence des bactéries pathogènes sont typiquement des régions où l'annotation est délicate, car elles peuvent concentrer plusieurs de ces difficultés dans des régions allant de 10 kb à 200 kb [Hacker & Kaper, 2000]. En effet, ces régions sont potentiellement soumises au transfert horizontal (CDS *atypiques*), au déclin (pseudogène), à la variabilité antigénique (duplication, répétition, inversion), à des régulations traductionnelles (*frameshift* authentique, ARN antisens), etc. Elles doivent donc être annotées avec beaucoup de soin si l'on veut avoir toutes les informations nécessaires à la compréhension des mécanismes biologiques de l'expression de ces gènes.

Le choix du modèle probabiliste de séquences d'ADN, est à replacer dans le contexte de 1999. En 1999, les programmes de prédiction de gènes procaryotes les plus précis utilisaient les modèles de Markov comme *GeneMark* (chaînes de Markov [Borodovsky & McIninch, 1993a]), *GLIMMER* (ICM [Salzberg *et al.*, 1998]) ou *GeneMark.hmm* (HSMM [Lukashin & Borodovsky, 1998]).
GeneMark (voir p. 93) modélise le non-codant et calcule les probabilités de codage le long des six phases grâce à la formule de Bayes appliquée à un fragment de séquence contenu dans une fenêtre glissante ($P(COD_m \mid F)$ avec $m \in \{-1, -2, -3, 0, +1, +2, +3\}$). A partir de ces vecteurs de

probabilités et de la liste des CDS de longueur maximale supérieure à un seuil de longueur, leur probabilité moyenne de codage (Pc) est calculée, puis elles sont sélectionnées selon un seuil de codage. Cet algorithme simple ignore le problème difficile de la gestion des CDS recouvrantes. *GeneMark* ne modélise pas l'hétérogénéité de composition des CDS ; cependant, il est possible d'exécuter k fois *GeneMark* avec chacune des k matrices de transition spécifiques d'un groupe de gènes. Il suffit ensuite de réconcilier les prédictions.
GeneMark.hmm 1.0 (voir p. 114) utilise l'hypothèse que les gènes procaryotes ne se chevauchent pas. Il utilise l'algorithme de Viterbi pour prédire les CDS, ce qui ne permet pas d'attribuer un score à chaque CDS. En revanche le modèle décrit le non-codant et l'hétérogénéité des CDS. Pour chacun des deux programmes *GeneMark* et *GeneMark.hmm*, il existe deux façons de construire une matrice typique et une matrice atypique. La première utilise le programme *GeneMark-Genesis* pour séparer un jeu de CDS prédites, en deux groupes (typique et atypique), grâce à la méthode de partitionnement automatique du K-means sur les valeurs RSCU (matrices pseudo-natives [Hayes & Borodovsky, 1998b]). La seconde construit une matrice typique native à partir de l'ensemble des CDS annotées dans les banques et une matrice atypique heuristique à partir d'une table d'usage des codons heuristique (voir p. 116 [Besemer & Borodovsky, 1999]). Les programmes *GeneMark* et *GeneMark.hmm* recherchent le RBS des CDS prédites dans une étape de post-traitement. Pour l'installation en local, des programmes développés par M. Borodovsky *et coll.*, seuls des exécutables payants sont disponibles.
GLIMMER 2.0 (voir p. 102) ne modélise ni le non-codant, ni l'hétérogénéité des CDS. Il évalue le potentiel de codage directement sur la CDS complète (pas de fenêtre glissante) par la probabilité du modèle connaissant la CDS ($P(COD_m \mid CDS)$, formule de Bayes). Il gère les problèmes de chevauchements dans des étapes de post-traitements en essayant de réajuster le codon d'initiation des CDS mais sans rechercher de RBS. Les sources sont accessibles librement sous la forme de modules facilement utilisables.

Nous avons choisi le modèle de chaînes de Markov plutôt que les IMM ou les HMM, car :

- C'est un modèle éprouvé, simple et souple ; il permet d'apprendre le non-codant avec des séquences natives, de prendre en compte l'hétérogénéité des CDS, il autorise les chevauchements.
- Les résultats obtenus grâce au calcul de la Pc des CDS semblent plus réalistes que ceux obtenus en calculant d'autres scores, ce qui explique en partie la grande précision du programme *GeneMark*. De plus, les courbes de probabilités de codage peuvent aider au choix du codon d'initiation.
- La modularité du programme *Prokov* dont nous disposions, nous permettait d'agencer les modules ou d'en développer de nouveaux selon notre expertise de bioinformaticiens, d'annotateurs et de biologistes.

7.2 Importance de l'utilisation de matrices de transition adaptées aux génomes

La prédiction de gènes sur une séquence d'ADN par chaînes de Markov nécessite d'avoir préalablement estimé les paramètres du modèle, c'est-à-dire d'avoir construit une matrice de transition à partir d'un jeu de séquences d'apprentissage. Les exemples de la figure 7.2 p. 204 permettent de visualiser les courbes de probabilités de codage d'un même fragment d'ADN de *B. subtilis* avec quatre matrices différentes. Dans l'exemple A, la matrice utilisée est construite à partir des gènes de classe I de *B. subtilis*. Elle prédit correctement les CDS de cette classe (ellipse bleue). En revanche, dans l'exemple C, la matrice construite à partir des gènes de classe I de *M. tuberculosis* H37Rv prédit mal les CDS de classe I de *B. subtilis* (ellipse bleue ; on observe des valeurs de probabilité de codage plus faibles). Ces résultats indiquent clairement que l'usage du code est très différent entre *B. subtilis* et *M. tuberculosis* H37Rv. La première règle qui se dégage est que *la matrice d'une espèce 1 ne peut être utilisée pour prédire les gènes d'une espèce 2 (*i.e. *il faut avoir un modèle de gènes spécifique de l'organisme étudié).* En revanche, on peut utiliser la matrice d'une souche 1 pour prédire les gènes d'une souche 2 de la même espèce (souches de *N. meningitidis*, d'*H. pylori*, de *M. tuberculosis*). Comme toutes les règles, elles comportent des exceptions. Par exemple, les génomes des deux espèces *Y. pestis* CO92 et *Y. pseudotuberculosis* présentent plus de 90% d'identité [Achtman *et al.*, 1999]. En conséquence, on peut utiliser la même matrice pour prédire les gènes de ces deux espèces (voir p. 48). A l'inverse, les deux souches *E. coli* K-12 et *E. coli* O157:H7 EDL933 ne sont proches qu'à 75% [Perna *et al.*, 2001]. En conséquence, il est plus raisonnable de ne pas utiliser la même matrice pour prédire les gènes de ces deux souches, même si elles sont de la même espèce. En pratique, pour décider si la matrice d'un génome 1 peut être utilisée sur un génome 2, il vaut donc mieux se fier au pourcentage d'identité nucléique entre les deux génomes qu'au système de classification des espèces bactériennes.

Dans l'exemple A, la matrice utilisée prédit mal les CDS de la classe III (ellipse rouge). En revanche dans l'exemple B, la matrice construite à partir des gènes de classe III de *B. subtilis*, prédit correctement les CDS de cette classe (ellipse rouge). Ces résultats indiquent clairement qu'à l'intérieur même d'un génome, il est possible de rencontrer des CDS n'ayant pas le même usage du code. Une des hypothèses pouvant expliquer le caractère atypique de certaines régions génomiques est celle d'un transfert horizontal suffisamment récent pour que l'usage des codons synonymes des gènes transférés soit encore très différent de celui des gènes natifs de la bactérie. Par conséquent, la deuxième règle est que *pour une espèce A, plusieurs matrices doivent être construites afin de prendre en compte l'hétérogénéité de composition dans les CDS* (*i.e.* dans l'idéal, il faut avoir autant de modèles de gènes qu'il y a de classes de gènes différentes dans leur usage des codons synonymes). Evidemment, dans l'exemple D, la matrice construite à partir des gènes de classe III de *M. tuberculosis* H37Rv prédit mal les CDS de classe III de *B. subtilis* (ellipse rouge ; on observe de nombreux pics de probabilités de codage parasites).

Les figures 7.2 E et F p. 204 mettent en évidence l'intérêt d'apprendre réellement le non-codant

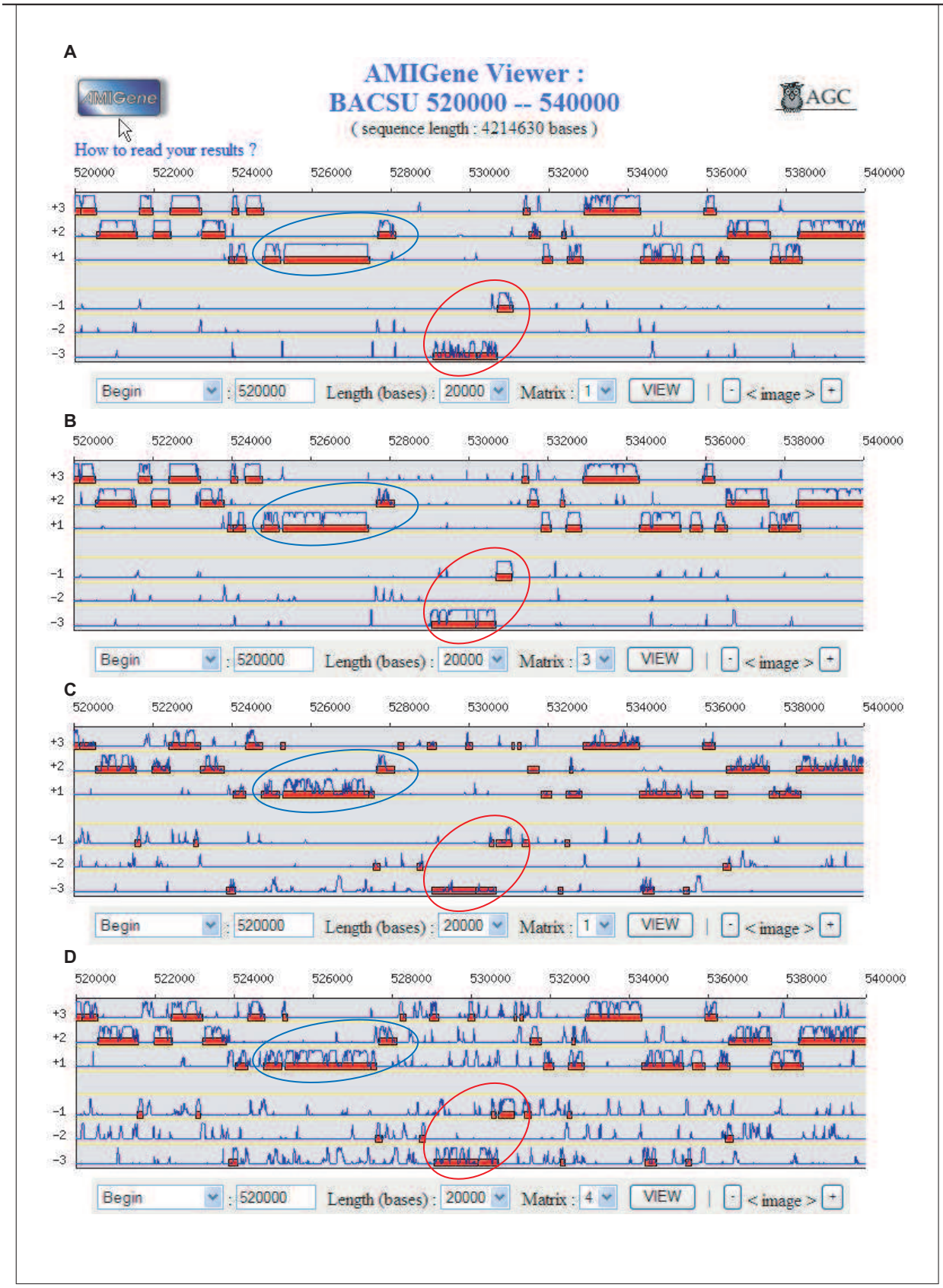


FIG. 7.2 – Utilisation de matrices de transition adaptées aux génomes

Visualisation des six phases de la région 520 - 540 kb du chromosome de *B. subtilis* où les CDS ont été prédites par le serveur Web *AMIGene*. Les rectangles rouges représentent les CDS prédite par *AMIGene* de longueur maximale et les courbes bleues représentent les valeurs de probabilité de codage calculée par Prokov_curve. Cette figure met en évidence l'importance de construire plusieurs matrices (ou modèles) par organisme.

A) utilisation de la matrice des gènes de classe I de *B. subtilis* sur un fragment du génome *B. subtilis* contenant des gènes typiques (ellipse bleue) suivie d'une région contenant des gènes atypiques (ellipse rouge) qui pourrait être issue d'un mécanisme de transfert horizontal.

B) utilisation de la matrice des gènes de classe III de *B. subtilis* sur la même région qu'en A.

C) utilisation de la matrice des gènes de classe I de *M. tuberculosis* H37Rv sur la même région qu'en A.

D) utilisation de la matrice des gènes de classe IV de *M. tuberculosis* H37Rv sur la même région qu'en A.

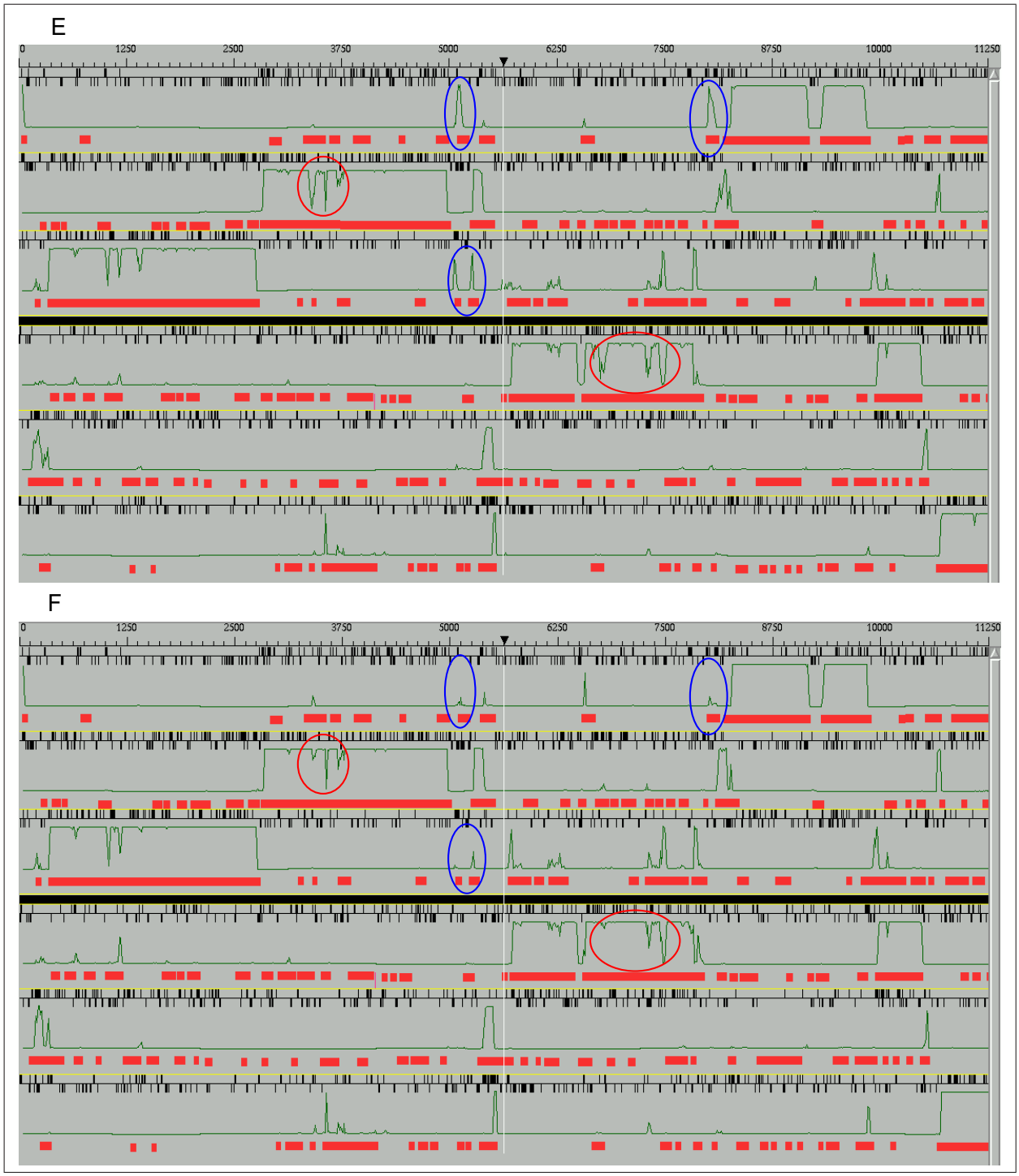


FIG. 7.2 – Utilisation du non-codant natif plutôt que du *shuffle* pour construire des matrices

Visualisation des six phases de la région 0-11 kb du chromosome d'*E. coli* K-12 dans le logiciel Imagene.

E) Utilisation de la pré-matrice générale d'*E. coli* K-12 par *prokov_curve*. Par *général* on entend toute classes de gènes confondues. Il y a un seul jeu de séquences codantes (toutes les CDS annotées dans les banques pour cet organisme). L'apprentissage (*prokov_learn*) s'est fait à partir d'un jeu de codant et les paramètres du non-codant sont calculés en mélangeant les effectifs du jeu de codant (*shuffle*). Autrement dit, une matrice contient deux modèles, le modèle du codant et le modèle du non-codant ; dans le cas des pré-matrices, le non-codant est généré par *shuffle*.

F) Utilisation de la matrice générale d'*E. coli* K-12 par *prokov_curve*. Le codant est modélisé de la même manière qu'en A. Le non-codant est modélisé par un jeu de séquences non-codantes native de l'organisme étudié, ici *E. coli* K-12. Les ellipses bleues mettent en évidence une diminution des pics parasites dans le non-codant entre les figures A et B. Les ellipses rouges mettent en évidence une augmentation de la courbe de probabilité de codage dans le codant entre les figures A et B. Finalement, l'apprentissage du non-codant permet une meilleure adéquation entre les courbes et les CDS.

plutôt que de le simuler en mélangeant les effectifs du codant (*shuffle*). Les courbes de probabilité de codage d'une région d'*E. coli* K-12 ont été calculées en utilisant la prématrice générale d'*E. coli* K-12 (*shuffle* ; exemple E) ou en utilisant la matrice générale d'*E. coli* K-12 (jeu de séquences non-codantes natives d'*E. coli* K-12 ; exemple F). On observe qu'utiliser un jeu de séquences non-codantes permet d'améliorer la prédiction du non-codant (moins de pics parasites au niveau des ellipses bleues) et aussi celle du codant (courbe plus régulière au niveau des ellipses rouges). Ces résultats indiquent clairement que le non-codant a une composition en oligonucléotides qui n'est pas le fruit du hasard. En conséquence, la troisième règle est que *l'apprentissage du codant natif* vs *non-codant* shuffle *n'est pas suffisant pour construire des matrices de transition de qualité* (*i.e.* il faut apprendre le codant natif *vs* le non-codant natif).

Enfin, la dernière règle extraite des résultats de M. Borodovsky *et coll.* est que *l'utilisation de l'ordre de la matrice, k, le plus élevé possible en fonction de la quantité de séquences disponible dans le jeu d'apprentissage, améliore à la fois la sensibilité et la spécificité de la phase de prédiction de gènes* (annexe C p. 389 [Borodovsky & McIninch, 1993b, Azad & Borodovsky, 2004]).

La phase d'apprentissage est cruciale pour la précision (sensibilité et spécificité) de la phase de reconnaissance par chaînes de Markov. En résumé, pour améliorer significativement la qualité des matrices, il est important d'utiliser : (i) des jeux d'apprentissage de séquences natives du génome, (ii) des jeux de CDS homogènes dans leur usage des codons synonymes, (iii) un jeu de séquences non-codantes et (iv) un ordre de chaînes de Markov aussi élevé que possible (*e.g.* $M5$) [Borodovsky & McIninch, 1993b, Borodovsky *et al.*, 1995, Hayes & Borodovsky, 1998b, Médigue *et al.*, 2002, Azad & Borodovsky, 2004]. Ainsi, nous avons développé la stratégie *AMIMat* qui permet de construire des matrices transition de haute qualité, étape préalable à une prédiction de gènes par chaînes de Markov précise. La stratégie que nous proposons est inspirée de *GeneMark-Genesis* de M. Borodovsky *et coll.* (voir p. 98 [Hayes & Borodovsky, 1998b]), c'est-à-dire que nous construisons une matrice par classe de gènes homogènes dans leur usage des codons synonymes.

7.3 Description de la stratégie *AMIMat*

Les méthodes de prédiction de gènes dépendant d'un modèle d'ADN codant contiennent toutes une phase d'apprentissage et une phase de reconnaissance. La phase d'apprentissage peut être complètement automatique ; elle est alors transparente pour l'utilisateur. Elle peut aussi être manuelle ; l'utilisateur doit alors fournir des jeux de séquences au programme d'apprentissage. Enfin, elle peut être semi-automatique ; l'utilisateur doit alors choisir certains paramètres et veiller au bon déroulement de chacune des étapes. Nous avons opté pour cette dernière approche. En effet, la phase d'apprentissage est cruciale pour la phase de reconnaissance et nous avons vu que, d'une manière générale, la tendance de l'analyse des génomes est à l'interactivité avec l'utilisateur *via* des interfaces ergonomiques (voir p. 152). La stratégie *AMIMat* enchaîne donc des étapes automatiques simples avec des points de contrôles manuels. Elle permet d'explorer un génome par étape, chaque étape devant être validée. Ainsi, une étape peut être enrichie de connaissances issues de la validation de

l'étape précédente. La méthode est découpée en deux étapes classiques : (i) construction de pré-matrices et (ii) construction de matrices en utilisant les pré-matrices. Ce principe de ré-estimation successive des paramètres permet d'améliorer la qualité des paramètres stockés dans les matrices.

Cette stratégie d'apprentissage des séquences d'ADN, experte et semi-automatique permet de construire des matrices de transition des chaînes de Markov en tenant compte :

- De l'hétérogénéité de composition entre les génomes et à l'intérieur des génomes, c'est-à-dire que plusieurs matrices sont construites pour chaque génome à partir des classes de gènes du chromosome procaryote étudié (ces classes sont définies selon le critère d'usage différentiel des codons synonymes des gènes qui les constituent).
- Des connaissances disponibles, nous avons mis au point deux variantes : celle utilisée dans un projet d'annotation où seule la séquence du chromosome est disponible, et celle utilisée dans un projet de réannotation lorsqu'un jeu de CDS annotées est disponible, en plus du chromosome.

7.3.1 Briques de base d'*AMIMat*

Apprentissage, reconnaissance et post-traitements

Pour développer la stratégie *AMIMat*, nous nous sommes procurés et nous avons développé différents modules qui servent à l'apprentissage, à la reconnaissance et aux post-traitements des CDS d'un chromosome procaryote.

- *prokov_orf* est le module de recherche de CDS par signal de *Prokov*, un programme de prédiction de gènes par chaînes de Markov développé par nos collaborateurs de l'INRIAlpes (variante modulaire et libre de *GeneMark* ; Tab. 3.3 p. 91 ; Fig. 3.2 p. 97 ; annexe C p. 389 [Romanet, 2001]). Il recherche simplement, sur les six phases de la séquence d'ADN, les CDS définies du codon d'initiation de la traduction le plus en $5'$ (*Leftmost Start* (LS)) au premier codon de terminaison rencontré en phase en $3'$ (stop), et dont la longueur maximale est supérieure à un certain seuil (*e.g.* $LS\_L \geq 63$ pb, codon de terminaison inclus). Autrement dit, *prokov_orf* recherche les CDS d'une certaine longueur maximale (LS_CDS) sur les six phases d'une séquence nucléique. Chaque LS_CDS est alors définie par quatre paramètres : son identifiant, sa position de début (premier nucléotide du codon d'initiation si la CDS est en sens direct ou dernier nucléotide du codon de terminaison si la CDS est en sens inverse), sa position de fin et sa phase de lecture sur la séquence (+1, +2, +3, -1, -2 et -3). A partir de ces annotations, on peut déduire d'autres annotations comme la longueur et l'orientation (directe ou inverse). Deux modifications doivent être appliquées aux résultats de *prokov_orf* :

  1. La phase des CDS est calculée selon la même convention que celle utilisée par *prokov_curve* (annexe C.2 p. 393).
  2. La position du stop est modifiée afin de correspondre à la convention stop inclus au lieu de stop exclus.

– *prokov_learn* est le module de construction de matrice de probabilités de transition de *Prokov*. Il est fondé sur le modèle statistique des chaînes de Markov qui permet de modéliser les CDS et les séquences non-codantes qui composent la séquence chromosomique. Autrement dit, une matrice de transition contient les fréquences absolues en oligonucléotides de longueur $m+1$, pour le modèle codant (chaîne de Markov à transitions 3-périodiques Mm_3) et pour le modèle non-codant (Mm), calculées respectivement à partir d'un jeu d'apprentissage de CDS et d'un jeu d'apprentissage de séquences non-codantes (optionnel ; s'il n'y a pas de jeu de séquences non-codantes, alors ces fréquences sont simulées par *shuffle* des fréquences du codant). L'ordre m des modèles non-codant et codant, est choisi en fonction de la taille des jeux d'apprentissage, selon les équations respectives (annexe C p. 389 et voir p. 93) :

$$\begin{aligned} \sum N &\geq 30 * 4^{(m+1)} \\ \sum N &\geq 90 * 4^{(m+1)}. \end{aligned}$$

Par exemple, pour des modèles non-codant et codant d'ordre 5, la fréquence des hexamères sera estimée sur des jeux d'apprentissage dont la taille sera au minimum, de 122880 et 368640 pb respectivement.

– *prokov_curve* est le module de reconnaissance de phases codantes de *Prokov* qui utilise les matrices construites par *prokov_learn*. Le calcul des probabilités de codage le long des six phases de la séquence d'ADN est identique à celui du programme *GeneMark* (voir p. 93). *prokov_curve* permet, à partir d'un critère Bayesien, de prédire le modèle (codant ou non-codant) qui s'ajuste le mieux sur une fenêtre glissant le long des six phases de la séquence d'ADN (*e.g.* w = 96 pb et s = 12 pb). Ainsi, pour chacun des six cadres, *prokov_curve* calcule la probabilité de codage aux positions de la séquence correspondant au milieu de la fenêtre. Ces valeurs de probabilité sont contenues dans six vecteurs de la longueur de la séquence, et peuvent être représentées graphiquement par six courbes de probabilité de codage.
– *prokov_score* applique la formule de Bayes directement à une CDS pour calculer, par exemple, la probabilité du modèle codant en phase 1 connaissant la CDS. Le score calculé par *prokov_score* introduit un biais qui favorise les grandes CDS[1] au détriment des petites. Nous avons donc développé le module *compute_Pc* (post-traitement) qui calcule la probabilité moyenne de codage des CDS (la Pc est un réel compris entre 0 et 1), prédites par *prokov_orf* ou annotées dans les banques, en utilisant les résultats de *prokov_curve*. C'est au moment du calcul de la probabilité moyenne de codage d'une CDS qu'il est judicieux d'essayer de réajuster la position de son codon d'initiation (voir p. 241).
– *max_Pc* permet de combiner les Pc obtenues sur un chromosome avec différentes matrices. Pour chaque CDS, la meilleure Pc est conservée avec le numéro de la matrice correspondant.
– *filter_L_Pc* permet de filtrer un fichier de CDS en fonction de leur longueur et de leur Pc. Par exemple, on peut ne vouloir garder que les CDS dont $L > 300$ et $Pc \geq 0,4$.

[1] Plus une CDS est longue et plus sa probabilité d'être codante en phase +1 tend vers 1.

- *CDS2NCDS* permet de déduire les positions de début et de fin des séquences non-codantes à partir d'une liste de CDS et de la longueur du chromosome.
- *lst2fna* permet d'extraire les séquences nucléiques au format fasta à partir d'une liste de positions de début et de fin et la séquence chromosomique.

Ces huit modules *prokov_orf*, *prokov_learn*, *prokov_curve*, *compute_Pc*, *max_Pc*, *filter_L_Pc*, *CDS2NCDS*, *lst2fna* sont au cœur de nos stratégies de prédiction de CDS procaryotes (*AMIMat* et *AMIGene* ; voir p. 237) et de notre processus de réannotation des génomes procaryotes complets (voir p. 275).

Méthodes d'analyse multivariée

Il s'agit de partionner automatiquement un jeu de CDS d'un chromosome en fonction des coordonnées de l'AFC du tableau d'usage des codons synonymes des CDS (annexe D p. 399 et voir p. 121). L'AFC est utilisée comme une étape préalable à la classification pour deux raisons : pour son pouvoir descriptifs (mettre en avant des facteurs latents inattendus) et pour son pouvoir de filtrage (travailler éventuellement sur des coordonnées factorielles moins nombreuses que les variables de départ ; voir p. 143 [Lebart *et al.*, 2000]).

Nous utilisons le programme *AFCcodon* d'H. Chiapello, qui calcule les fréquences relatives des codons synonymes ($F_{RS}(abc)$ [Chiapello, 1999]), et la fonction K-means de la plate-forme d'analyse statistique, R [R Development Core Team, 2003, Hartigan & Wong, 1979]. Par ailleurs, nous utilisons aussi les tâches d'usage de codons (*RSCU*), d'AFC et de nuées dynamiques (*Ether*) du module d'analyse statistique *GenoBool* de la plate-forme d'exploration génomique *Genostar* (voir p. 152 [Durand *et al.*, 2003]).

Après avoir analysé les résultats de l'AFC sur plusieurs génomes bactériens, avec et sans les codons cystéines, nous avons décidé de supprimer les colonnes correspondant aux codons TGC et TGT de la cystéine pour toute analyse future, de façon à se soustraire de ce biais artificiel (voir p. 132 [Moszer, 1996, Chiapello, 1999]). On réalise donc l'AFC d'un tableau de RSCU de taille $n*57$, qui décrit, pour chaque CDS du jeu, la manière dont sont utilisés les 57 codons (L > 201 pb codon de terminaison inclus). Les coordonnées des gènes ou des codons sur les principaux axes révélés par l'AFC sont ordonnées suivant le pourcentage d'inertie décroissant porté par chacun de ces axes (valeur propre λ). On considère qu'un axe porte une information significative si $\lambda \geq 100/57 \approx 2\%$ (voir la règle de Kaiser p. 125).

La détermination du nombre d'axes de l'AFC à prendre en compte et du nombre de classes à demander lors du « partionnement », ainsi que le choix de la meilleure partition restent des problèmes difficiles à résoudre et nécessitent donc une expertise humaine (voir p. 125). En pratique, une solution simple consiste à réaliser la classification sur tous les axes portant plus de 2% d'information. Cela revient à travailler sur une quinzaine d'axes portant plus de 50% de l'inertie cumulée ; ce qui est un compromis raisonnable entre travailler avec seulement les deux premiers axes, et travailler avec les 57 axes, sans pour autant utiliser la règle de l'éboulis des valeurs propres (voir p. 125).

La technique de l'AFC a l'avantage de permettre une analyse duale : les lignes et les colonnes du tableau de départ sont interchangeables (voir p. 125). Par exemple, dans le cadre de l'analyse

de l'usage des codons synonymes, il est équivalent d'étudier la manière dont se répartissent les codons dans le nuage de gènes ou d'étudier la manière dont se répartissent les gènes dans le nuage de codons. Ainsi, on peut superposer les gènes et les codons sur le même graphique (*e.g.* selon les coordonnées définies par l'AFC sur les deux premiers axes). Dans notre cas, nous cherchons à établir des classes de gènes en fonction de l'usage des codons synonymes ; mais pour trouver le nombre de classes, il est plus facile d'analyser le nuage de codons car il comporte généralement beaucoup moins d'individus : on peut donc identifier chaque point du nuage.

Dans un premier temps, on s'intéresse au nuage de codons : (i) on observe la manière dont se répartissent les codons synonymes sur les deux premiers axes de l'AFC et (ii) on repère les codons qui ont le plus contribué à la formation de chacun de ces axes. Imaginons le cas d'école où le premier axe séparerait à une extrémité les gènes dont les codons se terminent préférentiellement par G ou C et à l'autre extrémité, ceux dont les codons se terminent préférentiellement par A ou T (biais base faible (A ou T) - base forte (G ou C) sur la troisième base des codons synonymes), et où le deuxième axe séparerait les gènes dont les codons se terminent préférentiellement par G ou A, de ceux dont les codons se terminent préférentiellement par C ou T (biais purine (G ou A) - pyrimidine (C ou T) sur la troisième base des codons synonymes). Ainsi, on observerait pour les quartets de la proline, de la thréonine et de l'alanine, par exemple, quatre groupes de codons synonymes se trouvant dans les quatre rectangles délimités par les deux premiers axes, respectivement : (CCG, ACG, GCG), (CCA, ACA, GCA), (CCC, ACC, GCC) et (CCT, ACT, GCT). On suppose que les gènes se répartissent aussi en quatre classes en fonction des codons synonymes qu'il utilisent préférentiellement.

Dans un second temps, on classe les gènes en quatre groupes à partir de leur coordonnées sur le nombre d'axes pris en compte.

Dans un troisième temps, on passe à l'analyse du nuage de gènes. On superpose sur les deux premiers axes de l'AFC, le nuage de codons et le nuage de gènes où chaque gène a une couleur différente en fonction de sa classe. Si les quatre classes de gènes se répartissent de façon cohérente avec les quatre classes de codons alors les techniques AFC–partionnement se valident mutuellement : la partition semble avoir réussi et sera confirmée lorsqu'une signification biologique pourra être attribuée à chacune des classes. Une classe de gènes n'englobe pas forcément une classe de codons ; par exemple, la classe des gènes précoce chez *M. tuberculosis* H37Rv est à cheval sur la classe des codons se terminant par G et sur celle des codons se terminant par T (Fig. 7.4 p. 218).

Si les résultats ne sont pas en adéquation alors il faut recommencer[2] la classification jusqu'à ce qu'une solution satisfaisante soit trouvée (au sens biologique). On est souvent obligé de procéder par tâtonnements (essais–erreurs) : en relançant plusieurs fois une méthode (*e.g.* nuées dynamiques par la méthode des noyaux) avec les mêmes paramètres ou avec des paramètres différents (*e.g.* taille des noyaux), mais aussi en changeant de méthode (*e.g.* nuées dynamiques par la méthode des centroïdes).

[2]Par exemple, lorsque les centres initiaux sont tirés au hasard, l'exécution de deux K-means n'aboutira pas forcément au même résultat.

Les résultats de la partition finalement choisie sont utilisés pour construire des matrices de transition spécifiques de l'usage des codons synonymes avec *prokov_learn*. Nous avons donc besoin d'un fichier de séquences nucléotidiques par groupe de gènes. Un ancien stagiaire de l'Atelier de Génomique Comparative, C. Devine, a modifié le programme du K-means pour qu'il génère en sortie k fichiers de CDS (un fichier par groupe).

Seul un expert est capable de répondre à la question : les k classes qui se dégagent suivant les principales tendances mises en évidence par l'AFC sont-elles correctement délimitées par le partionnement automatique ? Le travail de l'expert ne s'arrête pas là. Nous verrons comment procéder pour expliquer le rôle biologique de chacune de ces classes (*e.g.* gènes typiques, gènes issus d'un mécanisme de transfert horizontal, gènes fortement exprimés ; voir p. 220).

7.3.2 Enchaînement des modules dans la stratégie *AMIMat*

Deux variantes d'*AMIMat* ont été développées pour construire des matrices de transition : (i) le génome est déjà annoté (*i.e.* un jeu de CDS est disponible dans les banques de séquences) et (ii) le génome n'est pas encore annoté (*i.e.* seule la séquence du chromosome est disponible).

AMIMat pour l'annotation d'un nouveau génome

La stratégie de construction de matrices de probabilités de transition des chaînes de Markov pour l'annotation d'un nouveau génome est décrite dans la figure 7.3 A p. 214. La première phase d'*AMIMat* consiste à construire des prématrices de transition pour définir un jeu d'apprentissage de séquences d'ADN codantes d'un génome (flèches vertes et rouges). Les CDS suffisamment longues sont d'abord recherchées dans les six phases le chromosome. Le paramètre de longueur de *prokov_orf* est calculé à partir de l'équation de la droite de régression $y = 17x - 200$ (pour définir un premier jeu d'apprentissage de CDS). L'équation de cette droite a été définie simplement, à partir des trois couples de valeurs suivants : aux pourcentages en G+C du chromosome de 43%, 51% et 66% correspondent respectivement des longueurs minimales de CDS de 500, 700 et 900 pb (pour définir un premier jeu d'apprentissage des CDS). Ensuite, des analyses multivariées permettent d'établir k classes de gènes en fonction de l'usage des codons synonymes (*e.g.* « partionnement » de CDS par nuées dynamiques sur les coordonnées de l'AFC d'un tableau de *RSCU*). Les étapes RSCU–AFC–« partionnement » sont capitales : elles nécessitent l'expertise d'un bioanalyste et se déroulent actuellement dans le module *GenoBool* de la plate-forme *Genostar*. Puis, pour chaque classe de gènes du génome, une *prématrice* de transition est construite à partir du jeu de CDS correspondantes en simulant un jeu de séquences non-codantes par *shuffle* (*k prokov_learn*).

Dans une seconde phase (flèches bleues et rouges), nous combinons une recherche de CDS de longueur supérieure à 60 pb (*prokov_orf*) et une mesure du potentiel de codage avec chaque prématrice (*k prokov_curve*). La probabilité moyenne de codage des CDS est calculée à partir des courbes de probabilité de codage (*k compute_Pc*). Pour chaque CDS, la meilleure Pc et le numéro de la matrice correspondant sont conservées (*max_Pc*). Le jeu d'apprentissage des séquences d'ADN

codantes correspond à un jeu de CDS (L > 201 et Pc > 0,4 pour au moins une des prématrices). Ces CDS sont séparées, comme précédemment en classes homogènes dans leur usage des codons synonymes (RSCU–AFC–Ether du module *GenoBool*). Le jeu d'apprentissage des séquences d'ADN non-codantes est déduit à partir d'un jeu de CDS (L > 60 et Pc > 0,1 ; *CDS2NCDS*). La longueur minimum des séquences non-codantes est de 100 pb [Borodovsky *et al.*, 1995]. Une *matrice* de transition est finalement construite à partir de chaque classe de CDS et du jeu de séquences non-codantes natives.

AMIMat pour la réannotation d'un génome public

La stratégie de construction de matrices de probabilités de transition des chaînes de Markov dans le cadre de la réannotation d'un génome public est décrite dans la figure 7.3 B p. 214 et dans l'*Article II* p. 274.

Comme précédemment, la première phase d'*AMIMat* consiste à construire des *prématrices* de transition (flèches vertes et rouges). Le fichier des annotations est analysé et les annotations sont chargées jusqu'à la table [Compare_Annotation] de notre base relationnelle PkGDB (voir p. 179). Les CDS dont les bornes sont cohérentes (CA_check = '*validated*') sont extraites. Ensuite, des analyses multivariées permettent d'établir k classes de gènes en fonction de l'usage des codons synonymes. Puis, une *prématrice* de transition est construite à partir de chaque classe de gènes du génome en simulant un jeu de séquences non-codantes par *shuffle* (*k prokov_learn*).

La seconde phase (flèches bleues et rouges) commence par une vérification manuelle des CDS dont le statut de cohérence des bornes n'est pas à '*validated*', en utilisant l'interface de correction CompAnnotViewer (*prokov_orf*, *k prokov_curve* ; voir le CA_check p. 186). Le jeu d'apprentissage des séquences d'ADN codantes est alors constitué des CDS et fCDS dont les bornes sont cohérentes (CA_check à '*validated*' ou à '*checked*'). Ces CDS sont séparées comme précédemment en classes homogènes dans leur usage des codons synonymes. Enfin, le jeu d'apprentissage des séquences d'ADN non-codantes est déduit à partir du jeu de toutes les CDS et fCDS (*CDS2NCDS* L > 99 pb). Une *matrice* de transition est finalement construite pour chaque classe de CDS en utilisant le jeu de séquences non-codantes natives.

7.4 Exemples d'application de la stratégie *AMIMat*

C'est dans la continuité des travaux sur l'usage des codons synonymes chez *B. subtilis* et *E. coli* K-12 de C. Médigue *et coll.* et I. Moszer *et coll.* [Médigue *et al.*, 1991, Moszer *et al.*, 1999] que nous avons étudié les biais entre les différentes classes de gènes. Ces auteurs ont mis en évidence, par des méthodes d'analyses multivariées, l'existence de trois classes de gènes chez ces deux organismes : la classe I des gènes natifs, la classe II des gènes hautement exprimés (*Predicted Highly eXpressed* (PHX)) et la classe III des gènes AT_3 riches, qui pourraient potentiellement provenir de transferts horizontaux (*Horizontal Gene Transfer* (HGT)).

7.4.1 Caractéristiques des génomes

CDS	*B. subtilis*	*E. coli*	*M. tuberculosis*
number	4107	4274	3996
GC_1	0,525	0,592	0,681
GC_2	0,360	0,407	0,501
GC_3	0,447	0,561	0,798
alt_start %	43,44	42,09	56,08
ATG %	77,78	88,55	61,34
CTG %	0,15	1,06	0,13
GTG %	9,45	8,15	33,75
TTG %	12,62	2,24	4,79
TAA %	62,47	63,18	15,72
TAG %	14,32	7,35	29,51
TGA %	23,21	29,47	54,77

Tab. 7.1 – Caractéristiques des CDS

Afin de montrer l'intérêt de la stratégie *AMIMat*, nous l'avons appliquée à plusieurs génomes modèles choisis de manière à couvrir la fourchette de pourcentages en G+C observés. En ce qui concerne la variante d'*AMIMat* pour la réannotation, connaissant les problèmes d'annotations, nous tenions absolument à prendre en référence des annotations les plus fiables possibles.

1. Le génome de *B. subtilis* de 4,2 Mb est G+C pauvre (43,5% [Kunst *et al.*, 1997]). Les annotations que nous avons utilisées pour *B. subtilis* correspondent à la version R16.1 de la base Subtilist (26 Avril 2001 [Moszer *et al.*, 2002]). Elles sont pour l'instant disponibles au format EMBL-EBI[3]. La séquence du chromosome a changé (voir p. 295).
2. Le génome d'*E. coli* K-12 de 4,6 Mb est G+C moyen (50,6% [Blattner *et al.*, 1997]). Les annotations d'*E. coli* K-12 correspondent à la version 17 de la banque EcoGene (16 juin 2003 [Rudd, 2000]). Elles sont disponibles dans un format tabulé[4]. La séquence chromosomique n'a pas changé (NC_000913.fna). Nous utilisons aussi les annotations fonctionnelles de *GenProtEC* [Serres *et al.*, 2004].
3. Le génome de *M. tuberculosis* H37Rv de 4,4 Mb est G+C riche (65,6% [Cole *et al.*, 1998]). Les annotations de *M. tuberculosis* H37Rv correspondent à la version R4 de la base TubercuList (8 Juillet 2002 [Camus *et al.*, 2002]). J.-C. Camus nous les a fournies au format EMBL-EBI. La séquence du chromosome a changé (voir p. 295).

Toutes les annotations ont été analysées et stockées dans la base de données PkGDB (voir p. 175). A partir de ces annotations et de la séquence chromosomique, on peut extraire les jeux de CDS et de séquences non-codantes.

[3] `ftp://ftp.pasteur.fr/pub/GenomeDB/SubtiList/FlatFiles/`

[4] `http://bmb.med.miami.edu/EcoGene/EcoWeb/`

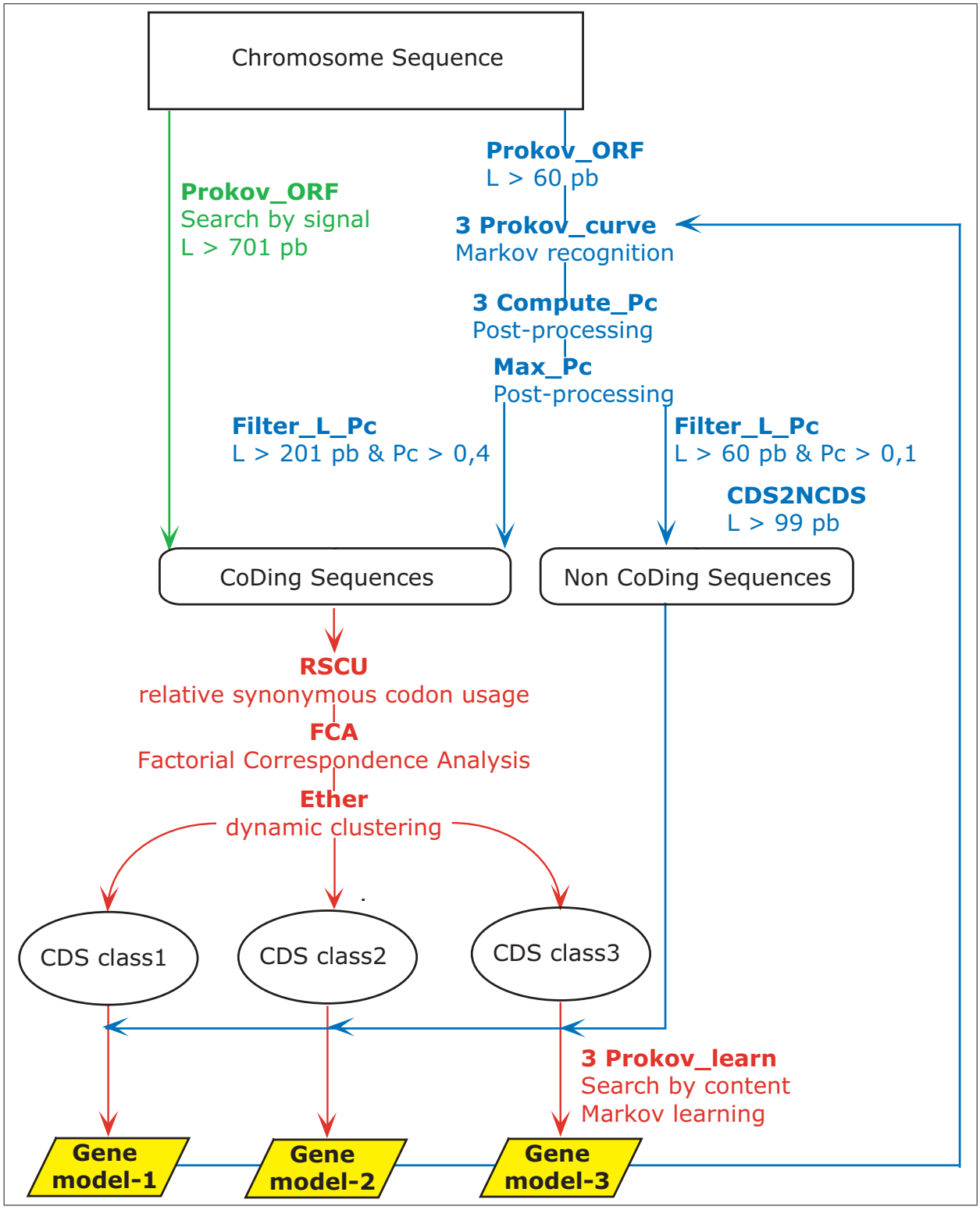


FIG. 7.3 – A) Variante *AMIMat* pour l'annotation d'un nouveau chromosome procaryote

Stratégie de construction de matrices de probabilités de transition des chaînes de Markov spécifiques des classes de gènes d'usage des codons synonymes (*e.g.* $k = 3$). Les étapes en vert sont spécifiques de la phase de construction de prématrices, les étapes en bleu sont spécifiques de la phase de construction de matrices et celles en rouge sont communes aux deux phases. Ces matrices sont dites pseudo-natives car elles utilisent un jeu de séquences codantes et non-codantes prédites automatiquement. L, longueur de la CDS codon de terminaison inclus et Pc, probabilité moyenne de codage.

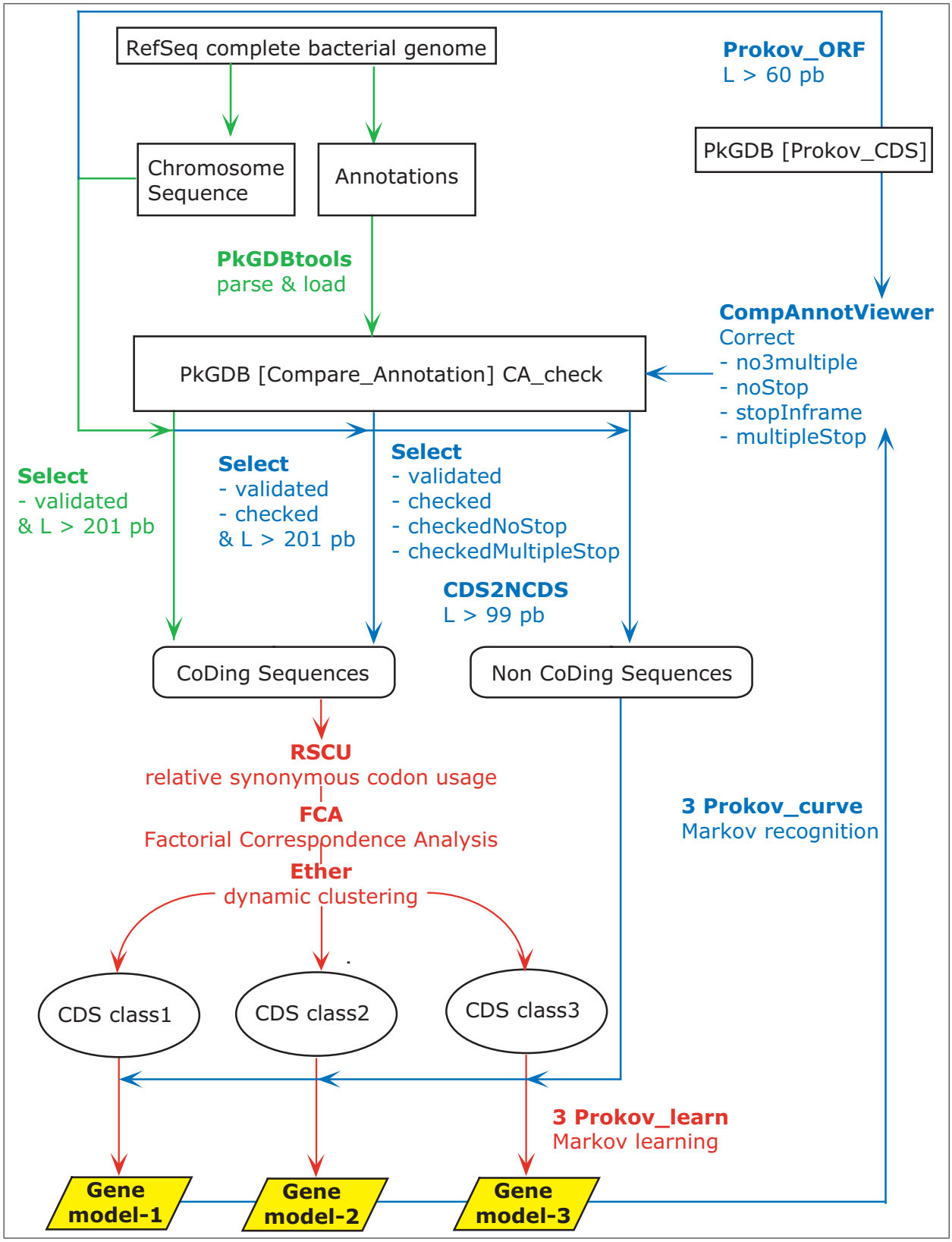


FIG. 7.3 – B) Variante *AMIMat* pour la réannotation d'un génome public

Ces matrices sont dites natives car elles utilisent un jeu de séquences codantes et non-codantes annotées (généralement validées manuellement).

En ce qui concerne la variante d'*AMIMat* pour l'annotation, nous avons choisi l'exemple de *P. luminescens* puisque nous avons participé à l'annotation de ce génome [Duchaud *et al.*, 2003] et que nous avons donc été confrontés à la construction de matrices avec, comme seule donnée, la séquence chromosomique. Son génome, de 5,7 Mb, est pauvre en G+C (42,8%).

Pour ces quatre génomes, la classification des gènes sur la base de leur usage des codons synonymes a conduit aux résultats suivants :

1. Nous confirmons les trois classes de gènes chez *B. subtilis* et *E. coli* K-12 [Médigue *et al.*, 1991, Moszer *et al.*, 1999].
2. Nous mettons en évidence quatre classes de gènes chez *M. tuberculosis* H37Rv et *P. luminescens*.

Ces résultats sont plus précisément décrits et validés dans les sous-sections suivantes.

7.4.2 Interprétation d'analyses multivariées

Les k classes de gènes définies pour les quatre génomes étudiés par la stratégie RSCU–AFC–« partionnement » sont présentés dans les figures 7.4 p. 218. Sur ces figures, un point représente un gène et une croix un codon dans le plan défini par les deux premiers axes de l'AFC (*i.e.* les plus informatifs, voir p. 125). Deux gènes dont les usages des codons synonymes sont similaires seront voisins sur le graphe. Les gènes et les codons sont respectivement coloriés en fonction de la classe de gènes et de la troisième base des codons. Ces graphes permettent de mettre en évidence des classes de gènes en fonction des biais de fréquence en troisième position des codons, dont les principaux sont les suivants :

1. biais entre bases faibles et bases fortes (AT_3 contre GC_3 ; atypique contre typique ; transfert horizontal ?),
2. biais purine–pyrimidine (GA_3 contre CT_3 ; RNY pyrimidine codons optimaux ; expressivité ?),
3. biais de réplication ou biais de brin (céto GT_3 contre amino CA_3 ; brin précoce contre brin tardif ; essentialité ?).

Chez les quatre génomes étudiés, le premier axe porte entre 9% (*M. tuberculosis* H37Rv) et 17% d'inertie (*E. coli* K-12) et le deuxième axe porte entre 7% (*P. luminescens*) et 9% (*B. subtilis*). Le nombre d'axes pris en compte lors du « partionnement », ainsi que d'autres paramètres, sont donnés dans la figure 7.4 F p. 218. Le premier axe ordonne les gènes et les codons selon le biais AT_3–GC_3, positionnant ainsi, à une extrémité de l'axe, les gènes de classe III relativement AT_3 riches par rapport au reste des gènes (Fig. 7.4 A p. 218). On remarque que les codons synonymes AGA et, dans une moindre mesure, AGG du sextet de l'arginine, sont des marqueurs des gènes de classe III qui permettent de l'identifier.

Le nuage de *M. tuberculosis* H37Rv a une forme allongée. L'axe 1 permet de séparer les gènes de classe I (GC_3 riches en vert) des gènes de classe III (AT_3 riches en orange ; Fig. 7.4 A p. 218). Les gènes de la classe I correspondraient aux gènes typiques, ce qui est cohérent avec la richesse

en G+C du génome. L'axe 2 permet de séparer les gènes de classe II en bleu des gènes de classe IV en rose selon un biais GT_3–CA_3 (Fig. 7.4 A p. 218). Les classes II et IV correspondraient donc respectivement aux gènes situés préférentiellement sur le brin précoce et sur le brin tardif [McInerney, 1998]. Il semble qu'aucune classe de gènes fortement exprimés n'ait été mise en évidence, ce qui est cohérent avec le fait que *M. tuberculosis* H37Rv est un organisme à croissance lente.

Chez *P. luminescens*, on retrouve un nuage de forme allongée découpé en quatre classes comme chez *M. tuberculosis* H37Rv, mais dans des proportions différentes (Fig. 7.4 F p. 218). La classe de gènes typiques correspondrait en fait à la somme des classes I (gènes situés préférentiellement sur le brin précoce) et IV (brin tardif). Cette fois-ci, la classe des gènes G+C riches correspondrait aux gènes fortement exprimés (classe II).

Chez *B. subtilis*, on observe un nuage de forme parabolique réparti en trois classes. L'axe 1 permet de séparer, selon un biais GC_3–AT_3, à une extrémité les gènes relativement[5] GC_3 riches (classe I) et à l'autre extrémité les gènes relativement AT_3 riches (classes II et III). L'axe 2 oppose la classe II à la classe III selon un biais C3 contre T3 (Fig. 7.4 A p. 218). Le deuxième axe permettrait donc d'affiner la séparation des gènes plutôt AT_3 riches : les gènes atypiques (HGT) qui utilisent préférentiellement des codons se terminant par A ou T, et les gènes hautement exprimés (PHX) qui utilisent préférentiellement des codons se terminant par C, ou T dans le cas des sextets (CGT de l'arginine, CTT de la leucine et TCT de la sérine ; Fig. 7.4 A p. 218). Ainsi, il se peut que parmi les gènes de classe II hautement exprimés, ceux qui possèdent en abscisse une valeur positive élevée proviennent de transferts horizontaux (gradient horizontal). Inversement, mais dans une moindre mesure, il se peut que les gènes de classes I et III qui possèdent en ordonnée une valeur positive importante soient faiblement exprimés (voire qu'ils soient des pseudogènes ; gradient vertical). C'est pourquoi il existerait, malgré le principe de l'AFC qui dit que les axes sont non corrélés entre eux, une certaine dépendance entre les deux premiers axes qui peut être à l'origine de l'effet Guttman (forme du nuage parabolique ou en fer à cheval).

Chez *E. coli* K-12, on retrouve un nuage de forme parabolique découpé en trois classes comme chez *B. subtilis*, mais les classes se forment le long des axes 1 et 2 selon un mécanisme différent. L'axe 1 permet de séparer à une extrémité les gènes de classe III (hautement atypiques) et à l'autre extrémité les gènes de classes II (hautement typiques) selon un biais AT_3–GC_3 (les gènes typiques de classe I se retrouvent au centre). L'axe 2 permettrait d'affiner la répartition des gènes des classe II et III en mettant en évidence un gradient d'expression selon un biais AT_3–GC_3. Dans la figure 7.4 A p. 218, les codons mis en évidence par les axes 1 et 2 sont différent bien que l'on observe globalement dans les deux cas un biais AT_3–GC_3. En effet, on a l'impression que le premier axe permettrait de séparer plutôt les codons se terminant par A en bleu, des codons se terminant par C en violet, alors que le deuxième axe séparerait les codons se terminant par G en rouge des codons se terminant par T en noir. Plus un gène GC_3 riche selon l'axe1 utiliserait le T en troisième position

[5] Le fait que les gènes typiques de classe I chez un génome A+T riche soient relativement G+C riches est tout à fait possible.

BACSU	**Axe 1**	**GC /**	**AT**			**Axe2**	**C/T**		
NNA	NNC	NNG	NNT	AA	NNA	NNC	NNG	NNT	AA
	ACC+	ACG+	ACT-	Threonine		AAC+		AAT-	Asparagine
CAA-		CAG+		Glutamine	AGA-	CGC+	AGG-	CGT+	Arginine
CCA-		CCG+	CCT-	Proline		CAC+		CAT-	Histidine
TTA-		CTG+		Leucine		TAC+		TAT-	Tyrosine
GCA-	GCC+	GCG+	GCT-	Alanine		TTC+		TTT-	Phenylalanine
	GGC+		GGT-	Glycine		GGC+	GGG-		Glycine
GTA-	GTC+	GTG+	GTT-	Valine					
ECOLI	**Axe 1**	**GC /**	**AT**			**Axe2**	**GC /**	**AT**	
	AAC-		AAT+	Asparagine					
ACA+	ACC-			Threonine	ACA+		ACG-	ACT+	Threonine
AGA+	CGC-	AGG+	CGT-	Arginine	AGA+	CGC-	AGG+	CGT-	Arginine
ATA+	ATC-			Isoleucine					
CAA+		CAG-		Glutamine					
	CAC-		CAT+	Histidine		CAC+		CAT-	Histidine
CCA+	CCC+	CCG-	CCT+	Proline	CCA+	CCC-	CCG-	CCT+	Proline
TTA+		CTG-	CTT+	Leucine	GCA+	GCC-	GCG-	GCT+	Alanine
GGA+	GGC-			Glycine			GGG-	GGT+	Glycine
	TAC-		TAT+	Tyrosine	GTA+		GTG-	GTT+	Valine
TCA+	TCC-			Serine			TCG-	TCT+	Serine
	TTC-		TTT+	Phenylalanine		TTC+		TTT-	Phenylalanine
MYCTU	**Axe 1**	**GC**	**AT**			**Axe2**	**GT/**	**CA**	
AAA-		AAG+		Lysine	AAA-		AAG+		Lysine
	AAC+		AAT-	Asparagine					
ATA-	ATC+			Isoleucine					
CAA-		CAG+		Glutamine	CAA-		CAG+		Glutamine
	CAC+		CAT-	Histidine					
CCA-			CCT-	Proline					
	GAC+		GAT-	Aspartate					
	TAC+		TAT-	Tyrosine		TAC-		TAT+	Tyrosine
	TTC+		TTT-	Phenylalanine					
PHOLU	**Axe 1**	**GC**	**AT**			**Axe2**	**GT/**	**CA**	
	AAC+		AAT-	Asparagine		AAC-		AAT+	Asparagine
ACA-	ACC+			Threonine					
AGA-	CGC+	AGG-	CGT+	Arginine	AGA-	CGC+		CGT+	Arginine
ATA-	ATC+			Isoleucine					
CAA-		CAG+		Glutamine	CAA-		CAG+		Glutamine
	CAC+		CAT-	Histidine		CAC-		CAT+	Histidine
		CCG+	CCT-	Proline	CCA-	CCC-	CCG+		Proline
	GAC+		GAT-	Aspartate	GAA-		GAG+		Glutamate
TTA-		CTG+		Leucine		GCC-	GCG+		Alanine
GGA-	GGC+			Glycine		GGC-	GGG+		Glycine
GTA-	GTC+	GTG+	GTT-	Valine	GTA-	GTC-	GTG+		Valine

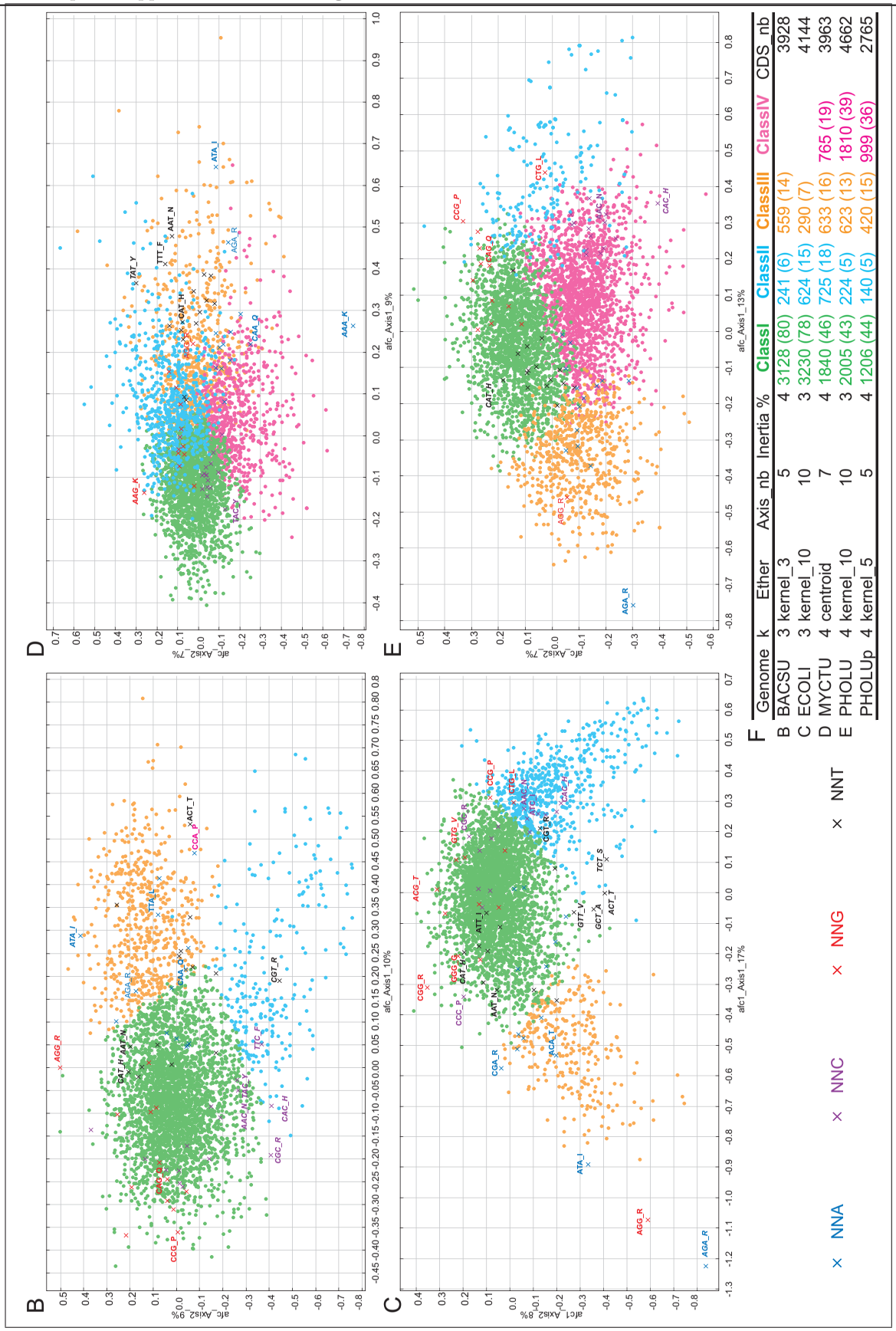
FIG. 7.4 – Séparation des codons synonymes en fonction des deux premiers axes de l'AFC

A) Ce tableau a été déduit à partir des valeurs contenues dans le tableau D.1 p. 405. Le signes plus et moins signifient respectivement que le codon a une coordonnée positive et négative sur l'axe. Les cases jaunes correspondent à des codons qui ont un comportement atypique par rapport aux autres codons qui ont fortement contribué à la formation de l'axe.

Sur les figures de la page recto, les codons qui ont le plus contribués à la formation de ces axes sont notés en gras pour l'axe 1 et en gras et en italique pour l'axe 2 (ct*100>380). Nous avons aussi repéré les codons AGA_R et AGG_R, marqueurs de l'usage des codons synonymes de la classe III. Il se peut que les valeurs négatives de la page verso correspondent à des valeurs positives dans la page recto et inversement. Il ne faut pas en tenir compte, l'orientation du nuage n'a pas d'importance dans l'AFC.

B) Parmi les trois classes de *B. subtilis*, l'axe 2 permet d'opposer le codon caractéristique des gènes de classe II GGC de la glycine au codon GGG caractéristique des gènes de classes I et III.

C) Parmi les trois classes d'*E. coli* K-12, le sextet de l'arginine est particulier : en effet, AGG est le seul codon se terminant par G caractéristique des gènes de classe III, CGA est le seul codon se terminant par A en classe I et CGT est plus proche des codons se terminant par C de la classe II que des codons se terminant par T des classes II et III. Le codon AGA est le premier codon qui contribue à la fois à l'axe 1 (ct 754) et à l'axe 2 (ct 749).



F

Genome	k	Ether	Axis_nb	Inertia %	ClassI	ClassII	ClassIII	ClassIV	CDS_nb
B BACSU	3	kernel_3	5	4	3128 (80)	241 (6)	559 (14)		3928
C ECOLI	3	kernel_10	10	3	3230 (78)	624 (15)	290 (7)		4144
D MYCTU	4	centroid	7	4	1840 (46)	725 (18)	633 (16)	765 (19)	3963
E PHOLU	4	kernel_10	10	3	2005 (43)	224 (5)	623 (13)	1810 (39)	4662
PHOLUp	4	kernel_5	5	4	1206 (44)	140 (5)	420 (15)	999 (36)	2765

FIG. 7.4 – Superposition du nuage de codons et des k classes de gènes sur les deux premiers axes de l'AFC

D) Parmi les quatre classes de *M. tuberculosis* H37Rv, le codon TAT a fortement contribué à la formation des deux axes (ct1 861 et ct2 792).

E) Parmi les quatre classes de *P. luminescens*, le codon CAC a fortement contribué à la formation des deux axes (ct1 498 et ct2 1132).

F) Statistiques de la méthode Ether de partionnement automatique par nuées dynamiques. La ligne PHOLUp correspond au jeu de CDS de *P. luminescens* définie par *prokov_orf* (L > 702 pb). Les autres jeux de CDS correspondent aux annotations des banques (L > 201).

des codons suivant l'axe2, et plus il serait fortement exprimé ; ce qui est cohérent avec le fait que les codons optimaux sont de la forme RNY [Shepherd, 1981]. Plus un gène AT_3 riche selon l'axe1 utiliserait T en troisième position des codons suivant l'axe2 et plus il serait fortement exprimé. Les gènes PHX d'*E. coli* K-12 utiliseraient donc préférentiellement les codons se terminant par C ou T. Ainsi, le biais AT_3–GC_3 mis en évidence par les deux premiers axes de l'AFC permet en fait de caractériser des propriétés biologiques différentes des gènes (*e.g.* gradient de composition en nucléotides, gradient d'expression) et pourrait être à l'origine de l'effet Guttman. Sur l'axe 1, le gradient de composition met en évidence des gènes AT_3 riches, GC_3 moyens et GC_3 riches. Le graphique de l'AFC présente la forme de deux oreilles de lapin, la première mettant en évidence le gradient d'expressivité des gènes AT_3 riches, et la seconde le gradient d'expression des gènes GC_3 riches. La classe majoritaire des gènes GC_3 moyens, à la jonction des gènes AT_3 riches et GC_3 riches, est une classe de gènes constitutivement exprimés : c'est la tête du lapin dans le graphe de l'AFC. Le placement du trio de l'isoleucine est facilement interprétable : ATT, ATC et ATA sont respectivement les codons préférés des gènes de classes I, II et III. En revanche, certains codons ont un placement surprenant sur le graphe de l'AFC (Fig. 7.4 p. 218). Le positionnement particulier des codons AGA et AGG de l'arginine n'est pas responsable de la forme de fer à cheval du nuage car si on retire le sextet de l'arginine de l'analyse, la forme persiste. Les codons positionnés avec les gènes de classe II se terminent par C ou T à l'exception du CTG du sextet la leucine. Ainsi, les codons des sextets peuvent avoir des positions particulières sur les axes de l'AFC, ce qui peut s'expliquer par le fait que les codons des sextets ne sont pas interchangeables [Diaz-Lazcoz *et al.*, 1995]. Le codon CCC du quartet de la proline est plus proche des codons se terminant par T de la classe I que des codons se terminant par C de la classe I. Des mécanismes différents peuvent donc parfois aboutir à des résultats similaires : chez *B. subtilis* et *E. coli* K-12, les classes ne sont pas mises en évidence selon le même usage des codons synonymes, mais dans les deux cas il semble exister des gènes qui soient à la fois atypiques et fortement exprimés.

7.4.3 Validation experte d'*AMIMat*

Avant d'utiliser la stratégie *AMIGene* (voir p. 237) pour annoter automatiquement un chromosome en utilisant de nouvelles matrices construites par *AMIMat*, un expert doit dans la mesure du possible : (i) expliquer le rôle biologique de chacune des classes de gènes d'un chromosome procaryote, définies en fonction de l'usage de codons synonymes et (ii) estimer la qualité des matrices à l'aider d'une interface graphique qui permet de visualiser les courbes de prédictions de codage pour chacune des matrices, notamment dans les régions atypiques du chromosome. Cette double vérification permet de valider définitivement le déroulement correct de la stratégie *AMIMat* complète.

Caractéristiques biologiques des classes de gènes

Informations complémentaires et corrélations Dans les quatre exemples présentés ci-dessus, il y a des points communs mais les interprétations sont toutes différentes. Par exemple, les génomes A+T riches de *B. subtilis* et *P. luminescens* présentent respectivement trois et quatre classes de gènes suivant l'usage des codons synonymes. Les génomes des bactéries à gram positif, *B. subtilis* et *M. tuberculosis* H37Rv, présentent respectivement trois et quatre classes de gènes. Les génomes des entérobactéries, *E. coli* K-12 et *P. luminescens*, présentent respectivement trois et quatre classes.

Pour faciliter l'interprétation biologique de la représentation de l'AFC, il est possible de colorer les gènes qui partagent une propriété biologique, comme l'expressivité, afin d'observer s'ils sont regroupés sur la représentation. De plus, des analyses statistiques de type régression linéaire permettent de corréler deux variables, par exemple, le GC_3 des CDS contre leur coordonnées sur l'axe1 de l'AFC. Les exemples de la figure 7.5 p. 222 nous permettent de donner des interprétations biologiques des classes de gènes qui valident le choix de k.

Pour les quatre génomes étudiés, l'axe 1 est toujours corrélé au taux de GC_3 (*i.e.* ordonnancement des gènes selon le GC_3), qui peut lui-même être lié ou non à l'expressivité. Par exemple, selon le premier axe chez *E. coli* K-12, les gènes hautement exprimés seraient relativement GC_3 riches alors que chez *B. subtilis*, ils seraient relativement AT_3 riches (en turquoise dans les figures 7.4 B et C p. 218 et 7.5 A et D p. 222). Plus précisément on remarque sur la figure 7.5 A p. 222, que les gènes de classe II de *B. subtilis* ont un GC_3 intermédiaire, puisque cette classe se retrouve à cheval sur les classes I et III.

Nous avons confirmé l'existence de trois classes de gènes suivant l'usage des codons synonymes pour les génomes de *B. subtilis* et *E. coli* K-12 (I–typique, II–HPX et III–HGT [Médigue *et al.*, 1991, Moszer *et al.*, 1999]). Ensuite, le rôle de deux premiers axes dans la formation de ces classes est expliqué :

- L'axe 1 ordonne les gènes selon leur composition en GC_3 et met principalement en évidence les gènes de classe III AT_3 riches.
- L'axe 2 affine ce premier tri, en ordonnant les gènes selon l'utilisation des codons synonymes optimaux pour la traduction (liée à l'expressivité) et en mettant principalement en évidence les gènes de classe II.

Les gènes qui n'utilisent ni les codons AT_3 ni les codons optimaux restent au centre pour former la classe I. Enfin, le modèle se complique si l'on ajoute la possibilité qu'un gène puisse être à la fois acquis par transfert horizontal et hautement exprimé, comme c'est par exemple le cas pour les gènes de toxine [Friis *et al.*, 2000]. Cette nouvelle possibilité peut être à l'origine d'un effet Guttman (voir p. 344).

Comme nous l'avons vu, il existe aussi des différences entre l'usage des codons synonymes d'*E. coli* K-12 et de *B. subtilis*, notamment, les codons préférés des gènes hautement exprimés sont plus GC_3 riches chez *E. coli* K-12 que chez *B. subtilis*. En effet, chez *E. coli* K-12, il existe une corrélation entre le GC_3 des gènes et les axes 1 et 2 : sur l'axe 1 les gènes des classes II et III

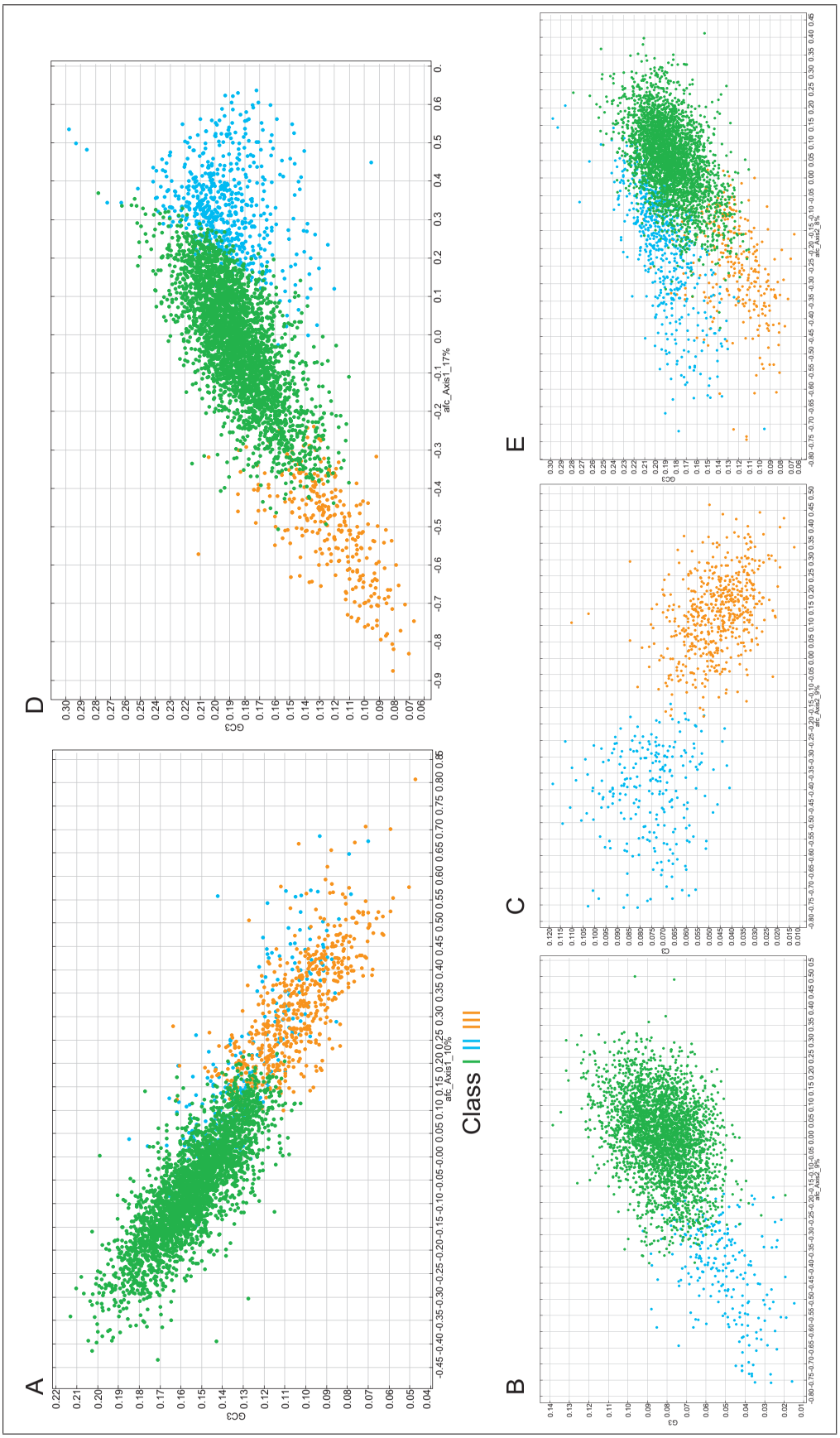


FIG. 7.5 – Informations complémentaires et corrélations

A) Taux de G+C en troisième position des codons de 3128 CDS de *B. subtilis* de classe I en vert, 241 CDS de classe II en bleu et 559 CDS de classe III en orange, en fonction de leurs coordonnées sur le premier axe de l'AFC.

B) Taux de G3 des CDS de *B. subtilis* de classe I et II en fonction de leurs coordonnées sur le deuxième axe de l'AFC.

C) Taux de C3 des CDS de *B. subtilis* de classe II et III en fonction de leurs coordonnées sur le deuxième axe de l'AFC.

D) Même graphe qu'en A) mais avec les 3230 CDS d'*E. coli* K-12 de classe I, les 624 CDS de classe II et les 290 CDS de classe III.

E) Taux de GC_3 des mêmes CDS qu'en C) en fonction de leurs coordonnées sur le deuxième axe de l'AFC.

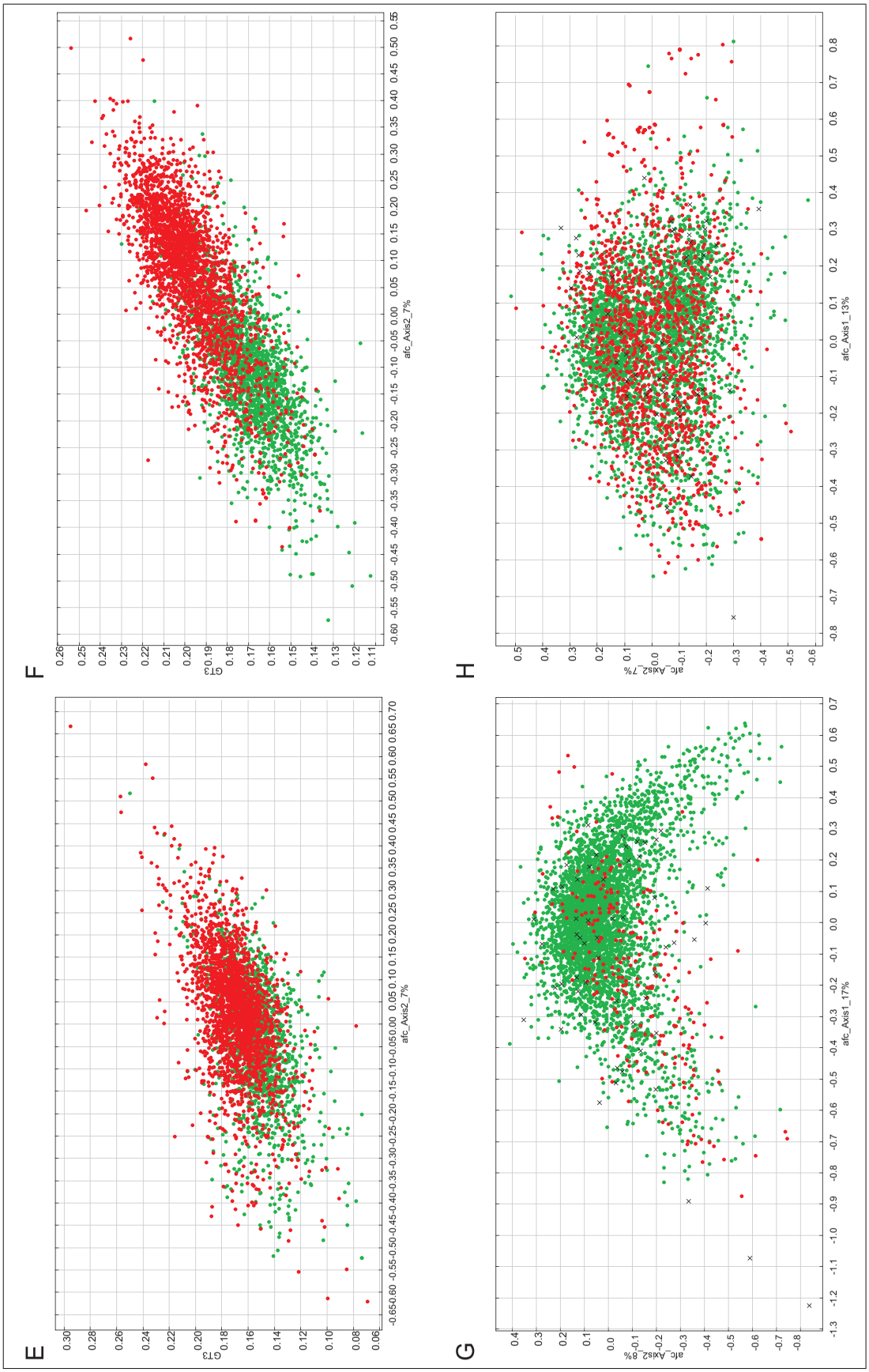


FIG. 7.5 – Informations complémentaires et corrélations

E) Taux de G+T en troisième position des codons de 3939 CDS de *M. tuberculosis* H37Rv en fonction de leurs coordonnées sur le second axe de l'AFC.

F) Même graphe qu'en E) mais avec les 4662 CDS de *P. luminescens*.

G) AFC sur les deux premiers axes des RSCU des 4144 CDS d'*E. coli* K-12, en rouge si elles sont d'origine extrachromosomique (type de produit *h* (ou 8) selon *GenProtEC* [Serres *et al.*, 2004]).

H) AFC sur les deux premiers axes des RSCU des 4662 CDS de PHOLU, en rouge si elles sont d'origine extrachromosomique (appartenance à un îlot génomique annoté [Duchaud *et al.*, 2003]).

se retrouvent à l'opposé (Fig. 7.5 D p. 222) ; sur l'axe 2, ils se retrouvent du même côté (double corrélation ou corrélation bifide Fig. 7.5 E p. 222), et, dans les deux cas, le GC_3 des gènes de classe II est supérieur à celui des gènes de classe III. Les gènes de classe II sont moins riches en GC_3 que les gènes de classe I, puisqu'ils utilisent rarement les codons se terminant par G, et les gènes de classe III sont moins riches en GC_3 que les gènes de classe II, puisqu'ils utilisent rarement les codons se terminant par G ou C. C'est pourquoi il n'y a plus qu'une seule corrélation lorsqu'on représente le G3 contre l'axe 2. Ainsi un biais de GC_3 peut révéler différentes propriétés biologiques.

Ce n'est pas le cas chez *B. subtilis* : on ne retrouve pas la double corrélation GC_3–axe 2. Les biais dans l'usage des codons synonymes sont moins marqués chez *B. subtilis* : on n'a que 19% de l'information dans le premier plan de l'AFC contre 25% chez *E. coli* K-12. Pour comprendre ce qui différencie les gènes de classe II des gènes de classes I et III suivant l'axe 2 nous avons procédé en deux étapes. La principale différence entre les gènes des classes I et II est que ces derniers utilisent rarement les codons se terminant par G (Fig. 7.5 B p. 222). La principale différence entre les gènes des classes II et III chez *B. subtilis* est que ces derniers utilisent rarement les codons se terminant par C (Fig. 7.5 C p. 222).

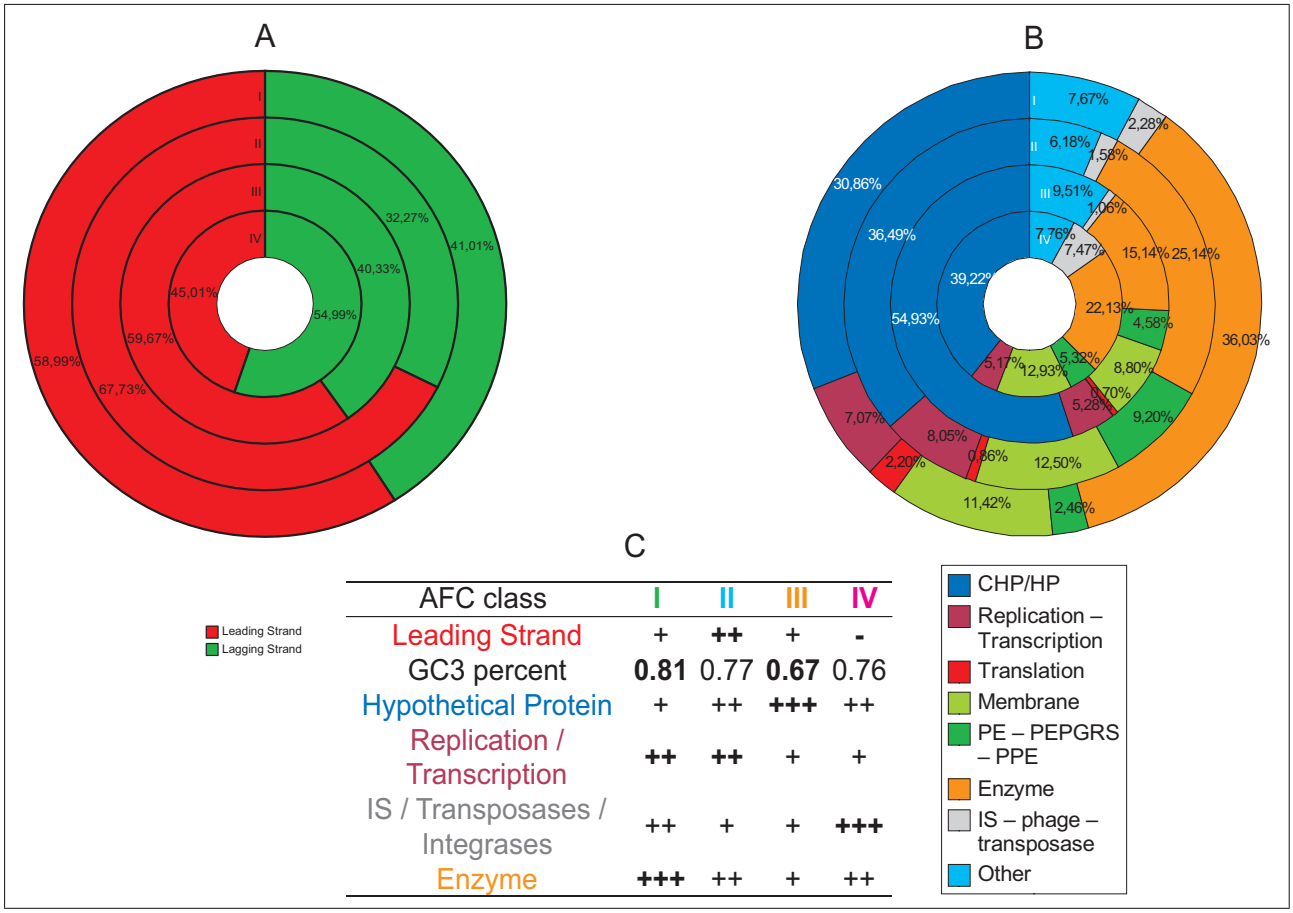
Chez *M. tuberculosis* H37Rv et *P. luminescens*, il s'agit de démontrer l'existence de quatre classes de gènes suivant l'usage des codons synonymes. Nous avons d'abord étudié l'usage des codons synonymes des gènes de *M. tuberculosis* H37Rv car ils sont particulièrement difficiles à prédire et nous avons suggéré qu'une partition en quatre classes semblait avoir un sens biologique [Cruveiller *et al.*, 2003a]. Puis, nous avons étudié l'usage des codons synonymes des gènes de *P. luminescens*, au moment de l'annotation de ce génome [Duchaud *et al.*, 2003]. La configuration du nuage des codons ressemblait à celle de *M. tuberculosis* H37Rv, et une partition en 4 classes semble effectivement adpatée :

- L'axe 1 permet de discriminer les gènes selon le biais bases faibles–bases fortes en troisième position des codons.
- L'axe 2 permet de séparer les gènes suivant leur orientation par rapport à l'origine de réplication.

C'est pourquoi on observe une corrélation GT_3 - axe 2 pour ces deux génomes, qui permet de séparer les gènes dont l'usage du code est principalement conditionné par le biais de réplication. Cette séparation devient évidente si l'on colorie les gènes en fonction de leur appartenance à l'un ou l'autre des brins réplicatifs (en rouge les gènes du brin précoce et en vert les gènes du brin tardif ; Fig. 7.5 E et F p. 222, et 7.6 A et C p. 225).

Il existe bien entendu, des différences dans l'usage des codons synonymes entre *M. tuberculosis* H37Rv et *P. luminescens*. Chez *M. tuberculosis* H37Rv, le biais d'expressivité n'est pas suffisamment fort pour que l'on puisse mettre en évidence une classe de gènes fortement exprimés (PHX). Nous avons identifié les gènes habituellement reconnus pour être fortement exprimés, plus précisément un échantillon de 26 gènes PHX codant des protéines ribosomiques[6] et des protéines de la machinerie

[6] *rps[ABCDEIJLM]* et *rpl[ABCDEIKLMNPQT]*.



AFC class	I	II	III	IV
Leading Strand	+	++	+	-
GC3 percent	**0.81**	0.77	**0.67**	0.76
Hypothetical Protein	+	++	+++	++
Replication / Transcription	++	++	+	+
IS / Transposases / Integrases	++	+	+	+++
Enzyme	+++	++	+	++

FIG. 7.6 – Analyse des 4 classes de gènes de *M. tuberculosis* H37Rv [Cruveiller *et al.*, 2003a]
A) Répartition des gènes sur le brin précoce et tardif dans les quatre classes de gènes de *M. tuberculosis* H37Rv. Les gènes situés préférentiellement sur les brins précoce et tardif appartiennent respectivement aux classes II et IV.
B) Répartition de diverses familles protéiques dans les quatre classes de gènes de *M. tuberculosis* H37Rv.
C) Tableau récapitulatif des propriétés des quatre classes de gènes de *M. tuberculosis* H37Rv.

de transcription[7] [Karlin *et al.*, 2001]. Ils sont regroupés avec les gènes de classe I GC_3 riches de *M. tuberculosis* H37Rv (Fig. 7.6 B p. 225). La classe I des gènes typiques a été défini comme telle car elle contient la majorité des gènes.

En revanche, chez *P. luminescens*, la classe II GC_3 riches contient les gènes habituellement reconnus pour être fortement exprimés et la classe des gènes typiques regroupe les classes I et IV qui sont les classes majoritaires. Nous avons testé un échantillon de vingt sept gènes PHX codant des protéines ribosomiques[8], la protéine chaperonne *dnaK* et des protéines de la machinerie de transcription[9]. Chez *B. subtilis* et *P. luminescens*, qui sont deux bactéries A+T riches, les gènes PHX n'auraient pas le même profil. Dans le cas de *B. subtilis*, ce seraient des gènes AT_3 riches selon l'axe1 qui utiliseraient préférentiellement le C en troisième position des codons suivant l'axe2, tandis que dans cas de *P. luminescens*, ce seraient des gènes GC_3 riches comme chez *E. coli* K-12 qui est aussi une entérobactérie.

Enfin, en utilisant les données disponibles, nous avons identifié les gènes HGT de *E. coli* K-12 [Serres *et al.*, 2004] et de *P. luminescens* [Duchaud *et al.*, 2003] pour confirmer que la classe III correspondait aux gènes issus de transferts horizontaux. L'origine extrachromosomique des gènes d'*E. coli* K-12 toucherait les classes I et III (et en proportion préférentiellement la classe III) mais pas la classe II (Fig. 7.5 G p. 222). En revanche, les gènes identifiés comme appartenant à des ilôts génomiques chez *P. luminescens* semblent répartis de manière uniforme dans les quatre classes (Fig. 7.5 H p. 222).

Nous pouvons aussi nous interroger sur un résultat a priori surprenant [Hayes & Borodovsky, 1998b] : si la classe de gènes A+T riches représentent les gènes issus de transferts horizontaux récents, alors pourquoi ne trouve-t-on pas de gènes de transfert issu d'organismes G+C riches ? Les séquences doivent posséder certaines propriétés structurales pour pouvoir être échangées facilement. Notamment, C. Friss *et coll.* ont remarqué que les régions A+T riches se *dépilent* plus facilement[10], qu'elles ont une courbure intrinsèque plus importante [11] et qu'elles sont moins flexibles[12] [Friis *et al.*, 2000]. Le chromosome bactérien est compacté à deux niveaux : la double hélice est liée à des protéines basiques[13] et est repliée en boucles d'environ 40 kb. Ce repliment met en jeu, à la base de la boucle, des interactions protéines - ADN (l'ensemble de la structure protéine–chromosome compactée est appelée le nucléoide). A. Perdersen *et coll.* expliquent que l'ADN en forme de rosette est compa-

[7] *rpo[CB]*, *fusA* et *tuf*.

[8] *rps[ABCDEIJLM]* et *rpl[ABCDEIKLMNPQT]*.

[9] *rpo[CB]*, *fusA* et *tsf*.

[10] *Stacking energy* : interaction entre bases adjacentes qui stabilise la structure tertiaire et favorise donc la compaction de l'ADN

[11] L'axe de la double hélice n'est pas une droite ; ne pas confondre enroulement et courbure. Par exemple, un plasmide superenroulé migre plus vite qu'un plasmide circulaire alors qu'un fragment courbé migre moins vite qu'un fragment non courbé.

[12] La flexibilité est mesurée comme la probabilité d'avoir le petit sillon à l'extérieur quand l'ADN est enroulé autour de protéines basiques.

[13] Sans les protéines basiques, la répulsion électrostatique de l'ADN due à sa charge empêcherait toute forme de compactage. L'interaction de l'ADN avec des protéines comme la protéine H-NS génère un ADN enroulé de manière plectonémique, ce qui est différent des nucléosomes chez les eucaryotes qui génèrent un enroulement de type solénoide.

tible avec l'idée que chaque boucle représente un domaine topologique isolé. Ils suggèrent que les régions hautement courbées (A+T riches) peuvent servir pour délimiter ces domaines. Sachant que les boucles du nucléoide sont des structures dynamiques et fluides, il est vraisemblable que l'ADN courbé reste au sommet des boucles la majeur partie du temps. A. Perdersen *et coll.* suggèrent que les régions hautement courbées (A+T riches) peuvent servir pour délimiter ces domaines : l'ADN courbé est vraisemblablement positionné au sommet d'une boucle une large partie du temps, empêchant la diffusion de supertours [Pedersen *et al.*, 2000]. En effet, un moyen d'absorber les supertours consiste à former des bulles de dénaturation, qui sont initiées dans des régions A+T riches et stabilisées par hybridation avec des fragments d'ADN simple brin en suspension [Strick *et al.*, 1998]. Ces trois travaux permettent d'imaginer que le transfert de gènes aurait lieu préférentiellement dans les régions courbées A+T riches au sommet des boucles du nucléoide, car ce sont des régions accessibles relativement à la base des boucles, et qui recombinent facilement (les régions A+T riches s'ouvrent et se *dépilent* plus facilement que les régions G+C riches). Ainsi, la composition des gènes HGT serait plus influencée par des propriétés structurales de l'ADN que par la composition du génome donneur. D'ailleurs une autre hypothèse expliquant la formation d'une classe III de gènes A+T riches est celle de la structure du chromosome bactérien [Daubin *et al.*, 2003b]. De plus, V. Daubin *et coll.* suggèrent que la composition nucléotidique des gènes transférés ne dépendent pas de celle du génome donneur [Daubin *et al.*, 2003a].

Par ailleurs, nous avons voulu identifier les gènes essentiels chez *E. coli* K-12 (base de données PEC[14]) et *B. subtilis* (base de données BSORF [Kobayashi *et al.*, 2003]) car un certain nombre de gènes codant des PHX sont aussi essentiels (*e.g.* les protéines ribosomiques, les composants des polymérases, les protéines chaperones ; [Rocha & Danchin, 2003]). L'essentialité toucherait les classes I et II (et en proportion, préférentiellement la classe II) mais pas la classe III.

Usage des codons synonymes Le tableau 7.2 p. 228 contient les fréquences relatives en nucléotides, en bases fortes (G+C), en bases puriques (G+A) et en bases céto (G+T) dans les différentes classes de CDS et dans le non-codant, calculé par le programme CodonW [Peden, 1999] pour les quatre génomes étudiés. La relative richesse en A+T des gènes de classe III est une caractéristique aussi partagée par les séquences non-codantes, ce qui pourrait expliquer pourquoi ces gènes sont plus difficiles à prédire par les méthodes fondées sur un modèle statistique d'ADN codant. Nous avons par ailleurs calculé les tables d'usage des codons synonymes pour les différentes classes de ces quatre génomes en utilisant CodonW (Fig. 7.7 p. 230).

Chez *M. tuberculosis* H37Rv, les gènes de classe I s'opposent aux gènes de classe III par le GC_3 (0,818 *vs* 0,684) et les gènes de classe II s'opposent aux gènes de classe IV par le GT_3 (0,525 *vs* 0,452). Les tables d'usage des codons synonymes (Fig. 7.7 p. 230) montrent aussi que les gènes de classe I sont les plus biaisés (*e.g.* leucine TTA 0,08 *vs* CTG 3,27) et que les gènes de classe III sont les moins biaisés (*e.g.* leucine TTA 0,24 *vs* CTG 2,11). Le codon de terminaison préféré est TGA. D'un point de vue intergénomique, le génome de *M. tuberculosis* H37Rv montre le biais

[14] `http://www.shigen.nig.ac.jp/ecoli/pec/`

	BS I	BS II	BS III	BS NCDS	EC I	EC II	EC III	EC NCDS	MT I	MT II	MT III	MT IV	MT NCDS	PL 1	PL II	PL III	PL IV	PL NCDS
CDS_nb	2627 (67)	573 (15)	728 (19)		2938 (71)	827 (20)	379 (9)		2288 (58)	686 (17)	275 (7)	690 (18)		1624 (35)	532 (11)	1115 (24)	1391 (30)	
Bio_FT	Typical	PHX	HGT		Typical	PHX	HGT		Typical	Le	HGT	La		Le Typical	PHX	HGT	La Typical	
Richness	G+C	C+T	A+T		G+C	C+T	A+T		G+C	G+T	A+T	C+A		G+T	G+C	A+T	C+A	
A	0,290	0,307	0,337	0,299	0,234	0,237	0,307	0,273	0,168	0,157	0,184	0,172	0,184	0,267	0,250	0,333	0,284	0,312
T	0,254	0,245	0,285	0,305	0,241	0,226	0,284	0,280	0,170	0,178	0,203	0,170	0,185	0,285	0,226	0,304	0,267	0,321
G	0,248	0,238	0,211	0,195	0,277	0,279	0,217	0,221	0,336	0,353	0,319	0,328	0,312	0,255	0,278	0,198	0,225	0,177
C	0,208	0,210	0,167	0,201	0,247	0,258	0,193	0,226	0,327	0,312	0,293	0,330	0,320	0,192	0,246	0,165	0,225	0,191
G+C	0,456	0,448	0,378	***0,396***	0,525	0,537	0,409	***0,447***	0,663	0,665	0,613	0,659	***0,632***	0,448	0,524	0,364	0,449	***0,368***
G1+C1	0,531	0,553	0,467		0,596	0,610	0,483		0,679	0,695	0,653	0,683		0,550	0,587	0,449	0,548	
G2+C2	0,363	0,366	0,336		0,413	0,401	0,369		0,492	0,521	0,501	0,521		0,389	0,420	0,348	0,386	
G3+C3	**0,475**	0,424	***0,331***		0,565	**0,601**	***0,376***		**0,818**	0,778	***0,684***	0,772		0,404	**0,565**	***0,294***	0,414	
G+A	0,538	0,545	0,548	0,493	0,512	0,517	0,524	0,494	0,504	0,509	0,504	0,500	0,496	0,522	0,528	0,531	0,508	0,488
G1+A1	0,632	0,669	0,641		0,593	0,630	0,613		0,612	0,618	0,591	0,611		0,603	0,593	0,622	0,600	
G2+A2	0,475	0,484	0,498		0,464	0,477	0,485		0,441	0,430	0,439	0,444		0,481	0,494	0,493	0,478	
G3+A3	0,507	***0,482***	0,506		0,478	***0,443***	0,474		0,459	0,480	0,481	0,445		0,482	0,498	0,479	0,447	
G+T	0,502	0,483	0,496	0,500	0,518	0,505	0,500	0,501	0,506	0,531	0,523	0,498	0,496	0,540	0,504	0,502	0,492	0,498
G1+T1	0,511	0,524	0,514		0,503	0,511	0,496		0,550	0,572	0,542	0,550		0,524	0,494	0,505	0,493	
G2+T2	0,464	0,448	0,450		0,491	0,472	0,464		0,486	0,496	0,496	0,492		0,481	0,465	0,461	0,467	
G3+T3	0,530	***0,476***	0,525		0,561	0,532	0,541		0,481	**0,525**	0,530	***0,452***		**0,615**	0,553	0,540	***0,515***	

TAB. 7.2 – Composition en nucléotides des k classes de gènes et du non-codant

Fréquences relatives en nucléotides, en bases fortes (G+C), en bases puriques (G+T) et en bases céto (G+T) dans les différentes classes de CDS et dans le non-codant chez *B. subtilis* (BS), *E. coli* K-12 (EC), *M. tuberculosis* H37Rv (MT) et *P. luminescens* (PL). Les codons en gras sont les codons fréquents et ceux en gras et italique sont rares. Les CDS utilisées dans ce tableau ont une longueur d'au moins 201 pb. BS : 3928 CDS (Ether centroid 4 axes) et 480457 pb pour le non-codant ; EC : 4144 CDS (Ether kernel-5 4 axes) et 540627 pb pour le non-codant ; MT : 3939 CDS (Ether centroid 4 axes) et 333373 pb pour le non-codant et PL : 4662 CDS (Ether centroid 5 axes) et 857456 pb pour le non-codant).

d'usage des codons synonymes le plus important entre les quatre génomes, par rapport à un usage équiprobable : l'écart à la moyenne du GC_3 des 3939 CDS est de 0,3 (TAB. 7.1 p. 213) et en classe I les codons se terminant par G ou C sont préférés (en gras dans le tableau 7.7 p. 230), à l'opposé des codons se terminant par A ou T, qui sont rares (en gras et en italique). En revanche, d'un point de vue intragénomique, *M. tuberculosis* H37Rv est le génome qui montre le moins de biais d'usage des codons synonymes entre les quatre classes de gènes, puisque la différence de GC_3 entre la classe I et III n'est que de 0,134.

Chez *P. luminescens*, les gènes de classe II s'opposent aux gènes de classe III par le GC_3 (0,565 *vs* 0,294) et les gènes de classe I s'opposent aux gènes de classe IV par le GT_3 (0,615 *vs* 0,515). On remarque, dans les tables d'usage des codons synonymes, que les gènes de classe III sont les plus biaisés (tous les acides aminés ont un codon rare (en gras et italique) et un codon préféré (en gras) ; FIG. 7.7 p. 230) et les gènes de classe IV sont les moins biaisés (*e.g.* la phénylalanine, l'asparagine, la cystéine n'ont ni codon rare ni codon préféré). Le codon de terminaison préféré est le TAA. D'un point de vue intergénomique, le génome de *P. luminescens* montre un biais d'usage des codons synonymes important : l'écart à la moyenne du GC_3 est de 0,1 et en classe III les codons se terminant par A ou T sont préférés, à l'opposé des codons se terminant par G ou C, qui sont rares. D'un point de vue intragénomique, *P. luminescens* est le génome qui montre le plus de biais d'usage des codons synonymes entre les quatre classes de gènes, puisque la différence de GC_3 entre la classe I et III est de 0,271.

Chez *B. subtilis*, les gènes de classe I s'opposent aux gènes de classe III par le GC_3 (0,475 *vs* 0,331) et les gènes de classe II s'opposent aux gènes de classe I et III par le GA_3 (0,482 *vs* 0,507 et 0,506). La table d'usage des codons synonymes des gènes de classe III est celle qui est la plus biaisée par rapport aux trois autres classes de *B. subtilis* (tous les acides aminés ont un codon rare

et un codon préféré, à l'exception de la cystéine). Les gènes de la classe I sont les moins biaisés (*e.g.* pour l'arginine en classe I l'écart va de CGA 0,59 à AGC 1,49 alors que pour l'arginine en classe II l'écart va de AGG 0,10 à CGT 2,07). Le codon de terminaison préféré est TAA.

Enfin, chez *E. coli* K-12, les gènes de classe II s'opposent aux gènes de classe III par le GC_3 (0,601 *vs* 0,376) et les gènes de classe II s'opposent aux gènes de classe I et III par le GA_3 (0,443 *vs* 0,478 et 0,474). On remarque, dans les tables d'usage des codons synonymes (Fig. 7.7 p. 230), que les gènes de classe II sont les plus biaisés (pour l'arginine en classe II, l'écart va de AGG 0,02 à CGT 3,18) et les gènes de la classe III sont les moins biaisés (*e.g.* pour l'arginine en classe III, l'écart va de AGG 0,57 à CGT 1,56). Le codon de terminaison préféré est le TAA.

Ainsi, certains de ces résultats corroborent des observations déjà décrites chez d'autres génomes :

- Selon [Shepherd, 1981, Borodovsky & McIninch, 1993a], les codons optimaux sont de la forme RNY.
- Selon [Smith & Eyre-Walker, 2001], les codons optimaux se terminent par G ou C et les codons sous-optimaux se terminent par A ou T.
- Selon [Hayes & Borodovsky, 1998b, Nicolas *et al.*, 2002], le codant est G+A riche ; les gènes atypiques et le non-codant sont A+T riches .
- Selon [Besemer & Borodovsky, 1999], les gènes de classe III sont proches du non-codant.

Qualité des matrices

La validation de la qualité des différentes matrices de transition d'un génome en fonction de l'usage des codons synonymes est facilitée par l'utilisation d'une interface cartographique (Fig. 7.8 p. 232). Chacun des quatre exemples de la figure 7.8 p. 232 présente des objets génomiques de PkGDB à travers l'interface *MaGe* et se découpe en trois cadres : (i) prédiction de codage avec la matrice I, (ii) avec la matrice III et (iii) résultat de synténie avec deux autres génomes. Les CDS en rouge sont communes aux annotations des banques et à la stratégie *AMIGene* (voir p. 237).

Dans l'exemple choisi pour *B. subtilis* (Fig. 7.8 A p. 232), la carte des synténies nous permet tout de suite de voir que nous sommes dans une région unique par rapport à *B. halodurans* et à *E. coli* K-12 (les orthologues colocalisés sont rares). La matrice III a permis à *AMIGene* de détecter une nouvelle CDS non annotée dans les banques (en bleu). La réalité de cette CDS est corroborée par la présence d'un orthologue chez *E. coli* K-12.

Dans l'exemple d'*E. coli* K-12 (Fig. 7.8 B p. 232), la région, choisie est commune avec *E. coli* O157:H7 EDL933 mais unique par rapport à *B. subtilis*. En $5'$ de cette région, on observe des décalages du cadre de lecture (CDS en haut de la phase) et en $3'$, on observe de courtes CDS difficiles à annoter (les CDS magenta n'ont pas été prédites par *AMIGene*). Pour valider que la classe III de *M. tuberculosis* H37Rv est mieux délimitée avec $k = 4$ qu'avec $k = 3$, nous avons construit les sept matrices de transition, puis nous avons calculé les courbes de probabilités de codage et les probabilités moyennes de codage des CDS. Ainsi, en comparant ces résultats sur une région atypique (*e.g.* annotée comme *low GC bias*), nous avons vérifié que nous améliorons significativement la prédiction des gènes de classe III lorsque $k = 4$. En effet, dans l'exemple C

BS I

Phe	**UUU**	**1,41**	Ser	UCU	1,10	Tyr	**UAU**	**1,32**	Cys	UGU	0,85
	UUC	***0,59***		UCC	0,84		***UAC***	***0,68***		UGC	1,15
Leu	UUA	1,03		**UCA**	**1,36**	TER	**UAA**	**1,77**	TER	UGA	0,80
	UUG	0,97		***UCG***	***0,69***		***UAG***	***0,44***	Trp	UGG	1,00
	CUU	**1,38**	Pro	CCU	1,04	His	**CAU**	**1,37**	Arg	CGU	0,93
	CUC	***0,74***		***CCC***	***0,39***		***CAC***	***0,63***		CGC	1,24
	CUA	***0,26***		***CCA***	***0,65***	Gln	CAA	0,93		***CGA***	***0,59***
	CUG	**1,63**		**CCG**	**1,92**		CAG	1,07		CGG	1,12
Ile	**AUU**	**1,51**	Thr	***ACU***	***0,51***	Asn	AAU	1,14	Ser	*AGU*	*0,55*
	AUC	1,14		***ACC***	***0,71***		AAC	0,86		**AGC**	**1,45**
	AUA	***0,35***		**ACA**	**1,58**	Lys	**AAA**	**1,36**	Arg	**AGA**	**1,49**
Met	AUG	1,00		ACG	1,20		***AAG***	***0,64***		***AGG***	***0,62***
Val	GUU	1,02	Ala	GCU	0,90	Asp	**GAU**	**1,26**	Gly	***GGU***	***0,63***
	GUC	1,13		GCC	0,91		***GAC***	***0,74***		**GGC**	**1,43**
	GUA	***0,69***		GCA	1,06	Glu	**GAA**	**1,32**		GGA	1,23
	GUG	1,16		GCG	1,13		***GAG***	***0,68***		***GGG***	***0,71***

EC I

Phe	UUU	1,24	Ser	***UCU***	***0,72***	Tyr	UAU	1,21	Cys	UGU	0,9
	UUC	0,76		UCC	0,81		UAC	0,79		UGC	1,1
Leu	UUA	0,83		***UCA***	***0,73***	TER	**UAA**	**1,82**	TER	UGA	0,95
	UUG	0,84		UCG	1,02		***UAG***	***0,23***	Trp	UGG	1
	CUU	***0,63***	Pro	CCU	0,64	His	CAU	1,23	Arg	**CGU**	**2,08**
	CUC	***0,64***		***CCC***	***0,58***		CAC	0,77		**CGC**	**2,5**
	CUA	***0,23***		CCA	0,76	Gln	***CAA***	***0,73***		***CGA***	***0,43***
	CUG	**2,82**		**CCG**	**2,02**		**CAG**	**1,27**		***CGG***	***0,7***
Ile	**AUU**	**1,62**	Thr	***ACU***	***0,58***	Asn	AAU	0,98	Ser	AGU	1
	AUC	1,18		ACC	1,68		AAC	1,02		**AGC**	**1,72**
	AUA	***0,2***		***ACA***	***0,54***	Lys	**AAA**	**1,53**	Arg	***AGA***	***0,18***
Met	AUG	1		ACG	1,2		***AAG***	***0,47***		***AGG***	***0,11***
Val	GUU	0,95	Ala	***GCU***	***0,58***	Asp	**GAU**	**1,31**	Gly	GGU	1,24
	GUC	0,91		GCC	1,13		***GAC***	***0,69***		**GGC**	**1,6**
	GUA	***0,58***		GCA	0,83	Glu	**GAA**	**1,35**		***GGA***	***0,48***
	GUG	**1,56**		**GCG**	**1,46**		***GAG***	***0,65***		***GGG***	***0,68***

BS II

Phe	UUU	0,98	Ser	**UCU**	**1,67**	Tyr	UAU	1,03	Cys	UGU	0,84
	UUC	1,02		***UCC***	***0,65***		UAC	0,97		UGC	1,16
Leu	UUA	1,20		**UCA**	**1,52**	TER	**UAA**	**2,45**	TER	***UGA***	***0,29***
	UUG	0,78		***UCG***	***0,30***		***UAG***	***0,27***	Trp	UGG	1,00
	CUU	**1,94**	Pro	**CCU**	**1,37**	His	CAU	1,04	Arg	**CGU**	**2,07**
	CUC	***0,60***		***CCC***	***0,13***		CAC	0,96		**CGC**	**1,89**
	CUA	***0,29***		CCA	0,92	Gln	CAA	1,19		***CGA***	***0,36***
	CUG	1,19		**CCG**	**1,57**		CAG	0,81		***CGG***	***0,34***
Ile	**AUU**	**1,51**	Thr	ACU	0,82	Asn	AAU	0,81	Ser	AGU	0,50
	AUC	**1,35**		***ACC***	***0,37***		AAC	1,19		**AGC**	**1,36**
	AUA	***0,14***		**ACA**	**1,92**	Lys	**AAA**	**1,56**	Arg	AGA	1,23
Met	AUG	1,00		ACG	0,90		***AAG***	***0,44***		***AGG***	***0,10***
Val	**GUU**	**1,31**	Ala	GCU	1,10	Asp	GAU	1,17	Gly	GGU	0,97
	GUC	0,91		***GCC***	***0,62***		GAC	0,83		**GGC**	**1,48**
	GUA	0,98		GCA	1,24	Glu	**GAA**	**1,48**		GGA	1,19
	GUG	0,81		GCG	1,04		***GAG***	***0,52***		***GGG***	***0,36***

EC II

Phe	UUU	0,77	Ser	**UCU**	**1,29**	Tyr	UAU	0,85	Cys	***UGU***	***0,75***
	UUC	1,23		**UCC**	**1,29**		UAC	1,15		**UGC**	**1,25**
Leu	***UUA***	***0,35***		**UCA**	***0,46***	TER	**UAA**	**2,3**	TER	***UGA***	***0,64***
	UUG	***0,47***		***UCG***	***0,76***		***UAG***	***0,05***	Trp	UGG	1
	CUU	***0,42***	Pro	***CCU***	***0,44***	His	CAU	0,77	Arg	**CGU**	**3,18**
	CUC	***0,59***		***CCC***	***0,18***		CAC	1,23		**CGC**	**2,48**
	CUA	***0,09***		**CCA**	***0,65***	Gln	***CAA***	***0,47***		***CGA***	***0,1***
	CUG	**4,08**		**CCG**	**2,74**		**CAG**	**1,53**		***CGG***	***0,16***
Ile	**AUU**	**1,26**	Thr	ACU	0,82	Asn	***AAU***	***0,49***	Ser	***AGU***	***0,46***
	AUC	**1,7**		**ACC**	**2,18**		AAC	1,51		**AGC**	**1,74**
	AUA	***0,04***		**ACA**	***0,25***	Lys	**AAA**	**1,57**	Arg	***AGA***	***0,04***
Met	AUG	1		ACG	0,76		***AAG***	***0,43***		***AGG***	***0,02***
Val	GUU	1,18	Ala	GCU	0,76	Asp	GAU	1,05	Gly	**GGU**	**1,66**
	GUC	***0,74***		GCC	0,95		GAC	0,95		**GGC**	**1,81**
	GUA	***0,66***		GCA	0,83	Glu	**GAA**	**1,46**		***GGA***	***0,17***
	GUG	**1,43**		**GCG**	**1,46**		***GAG***	***0,54***		***GGG***	***0,36***

BS III

Phe	**UUU**	**1,50**	Ser	**UCU**	**1,42**	Tyr	**UAU**	**1,45**	Cys	UGU	1,20
	UUC	***0,50***		***UCC***	***0,55***		***UAC***	***0,55***		UGC	0,80
Leu	**UUA**	**2,00**		**UCA**	**1,58**	TER	**UAA**	**1,76**	TER	***UGA***	***0,70***
	UUG	1,03		***UCG***	***0,45***		***UAG***	***0,54***	Trp	UGG	1,00
	CUU	**1,26**	Pro	**CCU**	**1,50**	His	**CAU**	**1,50**	Arg	CGU	0,94
	CUC	***0,43***		***CCC***	***0,41***		***CAC***	***0,50***		***CGC***	***0,58***
	CUA	***0,56***		CCA	1,25	Gln	**CAA**	**1,33**		CGA	0,84
	CUG	***0,72***		CCG	0,84		***CAG***	***0,67***		***CGG***	***0,57***
Ile	**AUU**	**1,54**	Thr	ACU	1,15	Asn	**AAU**	**1,36**	Ser	AGU	1,11
	AUC	***0,71***		***ACC***	***0,52***		***AAC***	***0,64***		AGC	0,90
	AUA	***0,75***		**ACA**	**1,71**	Lys	**AAA**	**1,42**	Arg	**AGA**	**2,30**
Met	AUG	1,00		***ACG***	***0,62***		***AAG***	***0,58***		AGG	0,77
Val	**GUU**	**1,54**	Ala	**GCU**	**1,36**	Asp	**GAU**	**1,47**	Gly	GGU	1,06
	GUC	***0,62***		***GCC***	***0,61***		***GAC***	***0,53***		GGC	0,82
	GUA	1,08		**GCA**	**1,44**	Glu	**GAA**	**1,42**		**GGA**	**1,46**
	GUG	0,76		***GCG***	***0,59***		***GAG***	***0,58***		***GGG***	***0,66***

EC III

Phe	**UUU**	**1,41**	Ser	UCU	1,16	Tyr	**UAU**	**1,42**	Cys	UGU	1,14
	UUC	***0,59***		***UCC***	***0,64***		***UAC***	***0,58***		UGC	0,86
Leu	**UUA**	**1,74**		**UCA**	**1,47**	TER	**UAA**	**1,81**	TER	UGA	0,76
	UUG	0,9		***UCG***	***0,54***		***UAG***	***0,43***	Trp	UGG	1
	CUU	1,16	Pro	**CCU**	**1,3**	His	**CAU**	**1,41**	Arg	**CGU**	**1,56**
	CUC	***0,52***		***CCC***	***0,65***		***CAC***	***0,59***		CGC	1,01
	CUA	***0,45***		**CCA**	**1,26**	Gln	CAA	1,06		CGA	0,87
	CUG	1,23		CCG	0,8		CAG	0,94		***CGG***	***0,6***
Ile	**AUU**	**1,48**	Thr	ACU	0,99	Asn	**AAU**	**1,38**	Ser	AGU	1,23
	AUC	***0,71***		ACC	0,93		***AAC***	***0,62***		AGC	0,96
	AUA	0,81		**ACA**	**1,27**	Lys	**AAA**	**1,5**	Arg	**AGA**	**1,39**
Met	AUG	1		ACG	0,8		***AAG***	***0,5***		***AGG***	***0,57***
Val	**GUU**	**1,53**	Ala	GCU	1,03	Asp	**GAU**	**1,47**	Gly	**GGU**	**1,42**
	GUC	0,78		GCC	0,83		***GAC***	***0,53***		GGC	0,88
	GUA	0,91		**GCA**	**1,34**	Glu	**GAA**	**1,4**		GGA	1
	GUG	0,78		GCG	0,79		***GAG***	***0,6***		***GGG***	***0,7***

FIG. 7.7 – Table d'usage des codons synonymes en fonction des k classes de gènes

Valeurs *RSCU* dans les trois classes de gènes de *B. subtilis* (BS) et *E. coli* K-12 (EC). Pour la légende, voir celle du tableau 7.2 p. 228.

MT I

Phe	***UUU***	***0,36***	Ser	***UCU***	***0,21***	Tyr	***UAU***	***0,47***	Cys	***UGU***	***0,46***
	UUC	***1,64***		***UCC***	***1,31***		***UAC***	***1,53***		***UGC***	***1,54***
Leu	***UUA***	***0,08***		***UCA***	***0,34***	TER	***UAA***	***0,47***	TER	***UGA***	***1,6***
	UUG	1,04		***UCG***	***2,2***		UAG	0,94	Trp	UGG	1
	CUU	***0,29***	Pro	***CCU***	***0,2***	His	***CAU***	***0,53***	Arg	***CGU***	***0,66***
	CUC	1,06		CCC	1,2		***CAC***	***1,47***		***CGC***	***2,44***
	CUA	***0,27***		***CCA***	***0,37***	Gln	***CAA***	***0,47***		***CGA***	***0,52***
	CUG	***3,27***		***CCG***	***2,23***		***CAG***	***1,53***		***CGG***	***2,05***
Ile	***AUU***	***0,41***	Thr	***ACU***	***0,21***	Asn	***AAU***	***0,36***	Ser	***AGU***	***0,34***
	AUC	***2,47***		***ACC***	***2,5***		***AAC***	***1,64***		***AGC***	***1,6***
	AUA	***0,11***		***ACA***	***0,27***	Lys	***AAA***	***0,4***	Arg	***AGA***	***0,08***
Met	AUG	1		ACG	1,02		***AAG***	***1,6***		***AGG***	***0,24***
Val	***GUU***	***0,33***	Ala	***GCU***	***0,29***	Asp	***GAU***	***0,49***	Gly	***GGU***	***0,74***
	GUC	***1,57***		***GCC***	***1,87***		***GAC***	***1,51***		***GGC***	***2,12***
	GUA	***0,2***		***GCA***	***0,36***	Glu	***GAA***	***0,66***		***GGA***	***0,38***
	GUG	***1,9***		***GCG***	***1,47***		***GAG***	***1,34***		GGG	0,77

MT II

Phe	***UUU***	***0,54***	Ser	***UCU***	***0,27***	Tyr	UAU	1,13	Cys	***UGU***	***0,58***
	UUC	***1,46***		UCC	1,13		UAC	0,87		***UGC***	***1,42***
Leu	***UUA***	***0,11***		***UCA***	***0,4***	TER	***UAA***	***0,43***	TER	***UGA***	***1,71***
	UUG	***1,36***		***UCG***	***2,2***		UAG	0,87	Trp	UGG	1
	CUU	***0,36***	Pro	***CCU***	***0,26***	His	***CAU***	***0,69***	Arg	***CGU***	***0,75***
	CUC	0,95		CCC	1,08		***CAC***	***1,31***		***CGC***	***2,06***
	CUA	***0,28***		***CCA***	***0,42***	Gln	***CAA***	***0,5***		***CGA***	***0,62***
	CUG	***2,94***		***CCG***	***2,25***		***CAG***	***1,5***		***CGG***	***2,17***
Ile	***AUU***	***0,53***	Thr	***ACU***	***0,3***	Asn	***AAU***	***0,51***	Ser	***AGU***	***0,47***
	AUC	***2,29***		***ACC***	***2,17***		***AAC***	***1,49***		***AGC***	***1,52***
	AUA	***0,18***		***ACA***	***0,31***	Lys	***AAA***	***0,54***	Arg	***AGA***	***0,12***
Met	AUG	1		ACG	1,22		***AAG***	***1,46***		***AGG***	***0,28***
Val	***GUU***	***0,41***	Ala	***GCU***	***0,37***	Asp	***GAU***	***0,67***	Gly	GGU	0,82
	GUC	***1,38***		***GCC***	***1,68***		***GAC***	***1,33***		***GGC***	***1,92***
	GUA	***0,23***		***GCA***	***0,38***	Glu	***GAA***	***0,69***		***GGA***	***0,41***
	GUG	***1,98***		***GCG***	***1,57***		***GAG***	***1,31***		GGG	0,85

PL I

Phe	**UUU**	**1,44**	Ser	**UCU**	**1,31**	Tyr	**UAU**	**1,55**	Cys	**UGU**	**1,38**
	UUC	***0,56***		***UCC***	***0,59***		***UAC***	***0,45***		***UGC***	***0,62***
Leu	**UUA**	**1,39**		UCA	1,05	TER	**UAA**	**1,67**	TER	UGA	0,96
	UUG	**1,71**		***UCG***	***0,59***		***UAG***	***0,38***	Trp	UGG	1
	CUU	0,76	Pro	CCU	1,18	His	**CAU**	**1,54**	Arg	**CGU**	**2,58**
	CUC	***0,33***		***CCC***	***0,44***		***CAC***	***0,46***		CGC	1,05
	CUA	***0,36***		CCA	1,08	Gln	CAA	0,94		***CGA***	***0,57***
	CUG	**1,45**		CCG	1,29		CAG	1,06		CGG	0,94
Ile	**AUU**	**1,79**	Thr	ACU	1,11	Asn	**AAU**	**1,44**	Ser	**AGU**	**1,53**
	AUC	0,76		ACC	1,03		***AAC***	***0,56***		AGC	0,93
	AUA	***0,44***		ACA	0,97	Lys	**AAA**	**1,47**	Arg	***AGA***	***0,55***
Met	AUG	1		ACG	0,88		***AAG***	***0,53***		*AGG*	*0,31*
Val	**GUU**	**1,48**	Ala	GCU	1,22	Asp	**GAU**	**1,65**	Gly	**GGU**	**1,74**
	GUC	***0,61***		***GCC***	***0,71***		***GAC***	***0,35***		GGC	0,84
	GUA	0,76		GCA	1,06	Glu	**GAA**	**1,35**		***GGA***	***0,65***
	GUG	1,16		GCG	1,01		***GAG***	***0,65***		GGG	0,78

PL II

Phe	**UUU**	**1,26**	Ser	***UCU***	***0,7***	Tyr	UAU	1,23	Cys	UGU	1,13
	UUC	0,74		UCC	0,77		UAC	0,77		UGC	0,87
Leu	UUA	0,92		UCA	0,88	TER	**UAA**	**1,42**	TER	UGA	1,24
	UUG	1,2		UCG	0,81		***UAG***	***0,34***	Trp	UGG	1
	CUU	***0,53***	Pro	***CCU***	***0,66***	His	CAU	1,14	Arg	**CGU**	**1,57**
	CUC	***0,57***		CCC	0,84		CAC	0,86		**CGC**	**1,57**
	CUA	***0,31***		***CCA***	***0,6***	Gln	CAA	0,77		***CGA***	***0,59***
	CUG	**2,48**		**CCG**	**1,9**		CAG	1,23		**CGG**	**1,55**
Ile	**AUU**	**1,25**	Thr	***ACU***	***0,6***	Asn	AAU	1,04	Ser	**AGU**	**1,26**
	AUC	1,19		**ACC**	**1,66**		AAC	0,96		**AGC**	**1,58**
	AUA	***0,55***		***ACA***	***0,65***	Lys	**AAA**	**1,39**	Arg	***AGA***	***0,39***
Met	AUG	1		ACG	1,1		***AAG***	***0,61***		***AGG***	***0,33***
Val	GUU	0,79	Ala	***GCU***	***0,71***	Asp	GAU	1,2	Gly	GGU	1,07
	GUC	0,95		**GCC**	**1,33**		GAC	0,8		**GGC**	**1,38**
	GUA	***0,6***		***GCA***	***0,7***	Glu	GAA	1,22		***GGA***	***0,6***
	GUG	**1,66**		**GCG**	**1,25**		GAG	0,78		GGG	0,95

MT III

Phe	UUU	0,82	Ser	***UCU***	***0,47***	Tyr	UAU	0,8	Cys	***UGU***	***0,67***
	UUC	1,18		UCC	1,09		UAC	1,2		***UGC***	***1,33***
Leu	***UUA***	***0,24***		***UCA***	***0,68***	TER	***UAA***	***0,58***	TER	***UGA***	***1,58***
	UUG	***1,38***		***UCG***	***1,78***		UAG	0,84	Trp	UGG	1
	CUU	0,69	Pro	***CCU***	***0,47***	His	CAU	0,85	Arg	CGU	0,88
	CUC	1,16		CCC	0,96		CAC	1,15		***CGC***	***1,82***
	CUA	***0,42***		***CCA***	***0,69***	Gln	CAA	0,76		CGA	0,91
	CUG	***2,11***		***CCG***	***1,88***		CAG	1,24		***CGG***	***1,71***
Ile	AUU	0,81	Thr	***ACU***	***0,5***	Asn	AAU	0,91	Ser	***AGU***	***0,63***
	AUC	***1,76***		ACC	1,67		AAC	1,09		***AGC***	***1,36***
	AUA	***0,43***		***ACA***	***0,56***	Lys	***AAA***	***0,67***	Arg	***AGA***	***0,24***
Met	AUG	1		***ACG***	***1,27***		***AAG***	***1,33***		***AGG***	***0,43***
Val	***GUU***	***0,69***	Ala	***GCU***	***0,57***	Asp	GAU	0,84	Gly	GGU	0,97
	GUC	***1,34***		GCC	1,4		GAC	1,16		***GGC***	***1,58***
	GUA	***0,36***		***GCA***	***0,55***	Glu	GAA	0,83		***GGA***	***0,61***
	GUG	***1,6***		***GCG***	***1,48***		GAG	1,17		GGG	0,85

MT IV

Phe	***UUU***	***0,43***	Ser	***UCU***	***0,28***	Tyr	***UAU***	***0,54***	Cys	***UGU***	***0,48***
	UUC	***1,57***		***UCC***	***1,28***		***UAC***	***1,46***		***UGC***	***1,52***
Leu	***UUA***	***0,12***		***UCA***	***0,46***	TER	***UAA***	***0,49***	TER	***UGA***	***1,77***
	UUG	1,01		***UCG***	***1,87***		***UAG***	***0,74***	Trp	UGG	1
	CUU	***0,39***	Pro	***CCU***	***0,27***	His	***CAU***	***0,58***	Arg	***CGU***	***0,68***
	CUC	1,2		***CCC***	***1,26***		***CAC***	***1,42***		***CGC***	***2,33***
	CUA	***0,38***		***CCA***	***0,55***	Gln	***CAA***	***0,7***		***CGA***	***0,71***
	CUG	***2,91***		***CCG***	***1,92***		***CAG***	***1,3***		***CGG***	***1,86***
Ile	***AUU***	***0,45***	Thr	***ACU***	***0,26***	Asn	***AAU***	***0,43***	Ser	***AGU***	***0,4***
	AUC	***2,35***		***ACC***	***2,36***		***AAC***	***1,57***		***AGC***	***1,71***
	AUA	***0,2***		***ACA***	***0,39***	Lys	AAA	1,07	Arg	***AGA***	***0,15***
Met	AUG	1		ACG	0,99		AAG	0,93		***AGG***	***0,28***
Val	***GUU***	***0,42***	Ala	***GCU***	***0,36***	Asp	***GAU***	***0,57***	Gly	***GGU***	***0,72***
	GUC	***1,61***		***GCC***	***1,84***		***GAC***	***1,43***		***GGC***	***2,14***
	GUA	***0,26***		***GCA***	***0,46***	Glu	GAA	0,82		***GGA***	***0,43***
	GUG	***1,7***		***GCG***	***1,33***		GAG	1,18		***GGG***	***0,71***

PL IV

Phe	UUU	1,18	Ser	UCU	1,23	Tyr	**UAU**	**1,34**	Cys	UGU	1,15
	UUC	0,82		UCC	0,88		***UAC***	***0,66***		UGC	0,85
Leu	**UUA**	**1,33**		UCA	1,13	TER	**UAA**	**1,91**	TER	UGA	0,79
	UUG	0,97		***UCG***	***0,36***		***UAG***	***0,3***	Trp	UGG	1
	CUU	0,79	Pro	CCU	1,11	His	CAU	1,23	Arg	**CGU**	**2,44**
	CUC	***0,55***		***CCC***	***0,68***		CAC	0,77		**CGC**	**1,65**
	CUA	***0,64***		**CCA**	**1,37**	Gln	CAA	1,23		***CGA***	***0,62***
	CUG	**1,73**		CCG	0,85		CAG	0,77		***CGG***	***0,61***
Ile	**AUU**	**1,52**	Thr	ACU	1,07	Asn	AAU	1,19	Ser	AGU	1,19
	AUC	1,07		**ACC**	**1,38**		AAC	0,81		AGC	1,21
	AUA	***0,41***		ACA	1,01	Lys	**AAA**	**1,61**	Arg	***AGA***	***0,52***
Met	AUG	1		***ACG***	***0,54***		***AAG***	***0,39***		***AGG***	***0,17***
Val	**GUU**	**1,49**	Ala	GCU	1,1	Asp	**GAU**	**1,48**	Gly	**GGU**	**1,68**
	GUC	0,92		GCC	1,05		***GAC***	***0,52***		GGC	1,22
	GUA	0,8		GCA	1,21	Glu	**GAA**	**1,54**		***GGA***	***0,63***
	GUG	0,78		***GCG***	***0,64***		***GAG***	***0,46***		***GGG***	***0,47***

PL III

Phe	**UUU**	**1,51**	Ser	**UCU**	**1,46**	Tyr	**UAU**	**1,56**	Cys	**UGU**	**1,38**
	UUC	***0,49***		***UCC***	***0,51***		***UAC***	***0,44***		***UGC***	***0,62***
Leu	**UUA**	**2,29**		**UCA**	**1,59**	TER	**UAA**	**1,77**	TER	UGA	0,77
	UUG	1,01		***UCG***	***0,4***		***UAG***	***0,46***	Trp	UGG	1
	CUU	1,02	Pro	**CCU**	**1,49**	His	**CAU**	**1,54**	Arg	**CGU**	**1,26**
	CUC	***0,38***		***CCC***	***0,52***		***CAC***	***0,46***		***CGC***	***0,67***
	CUA	***0,65***		**CCA**	**1,39**	Gln	**CAA**	**1,36**		CGA	0,82
	CUG	***0,64***		***CCG***	***0,6***		***CAG***	***0,64***		***CGG***	***0,46***
Ile	**AUU**	**1,49**	Thr	ACU	1,21	Asn	**AAU**	**1,52**	Ser	**AGU**	**1,34**
	AUC	***0,52***		***ACC***	***0,75***		***AAC***	***0,48***		***AGC***	***0,7***
	AUA	0,99		**ACA**	**1,49**	Lys	**AAA**	**1,53**	Arg	**AGA**	**2,1**
Met	AUG	1		***ACG***	***0,54***		***AAG***	***0,47***		***AGG***	***0,7***
Val	**GUU**	**1,68**	Ala	**GCU**	**1,37**	Asp	**GAU**	**1,64**	Gly	**GGU**	**1,5**
	GUC	***0,59***		***GCC***	***0,65***		***GAC***	***0,36***		***GGC***	***0,69***
	GUA	1,08		**GCA**	**1,39**	Glu	**GAA**	**1,47**		GGA	1,2
	GUG	***0,65***		***GCG***	***0,59***		***GAG***	***0,53***		***GGG***	***0,61***

FIG. 7.7 – Table d'usage des codons synonymes en fonction des k classes de gènes
Valeurs *RSCU* dans les quatre classes de gènes de *M. tuberculosis* H37Rv (MT) et *P. luminescens* (PL).

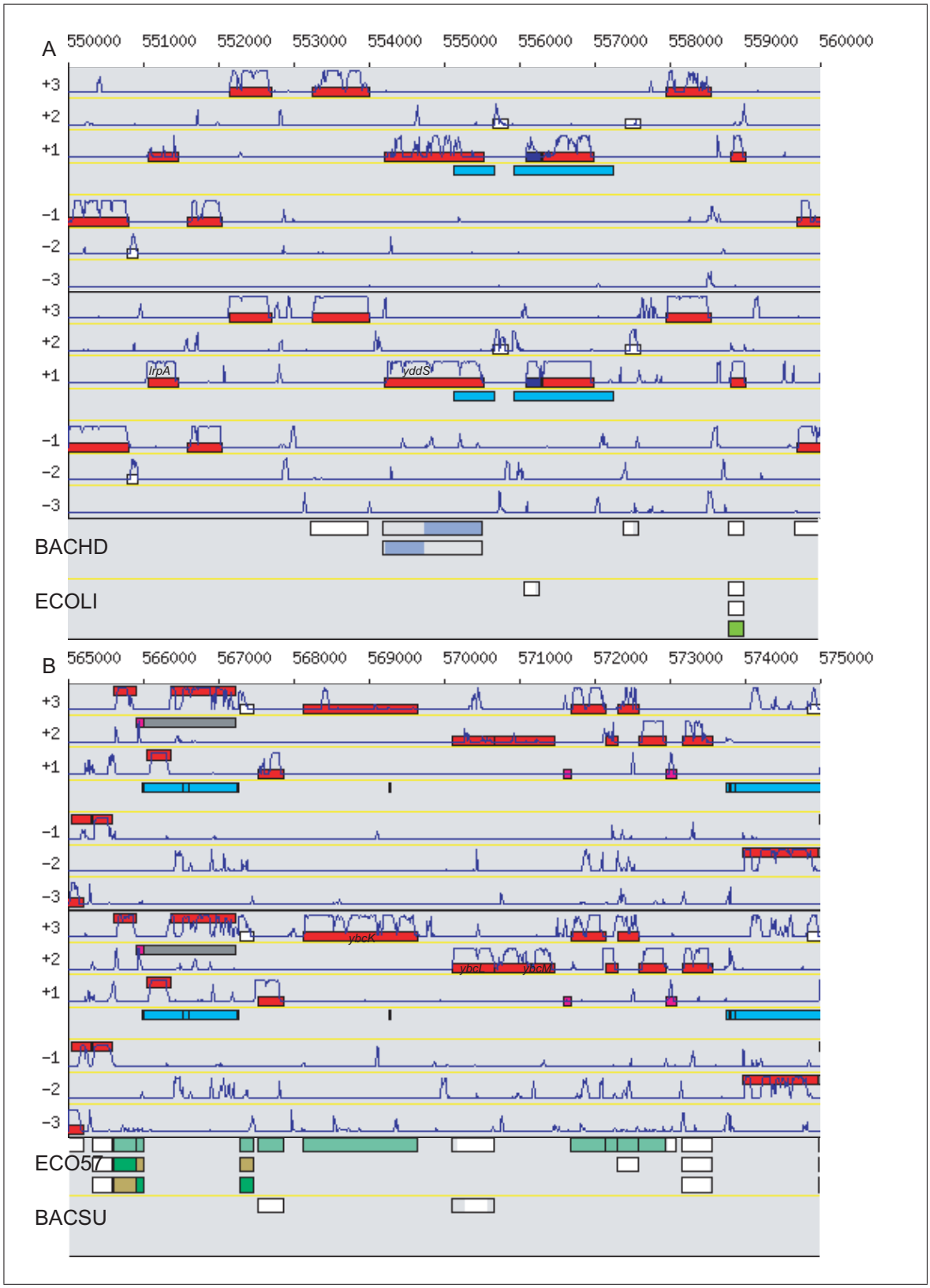


FIG. 7.8 – Validation des matrices de transition en fonction des k classes de gènes

Ces régions contiennent des gènes caractéristiques de l'usage des codons synonymes atypiques, qui montrent une faible probabilité de codage avec la matrice de transition des gènes de classe I et qui sont mieux prédits par la matrice III.

A) Les gènes de *B. subtilis*, *lrpA* et *yddS* codent respectivement un régulateur transcriptionnel de la famille Lrp–Asn et un polypeptide de fonction inconnue. Cependant, YddS est similaire à des séquences dont la fonction est impliquée dans le transport de métabolites (dont la résistance à l'antibiotique tetracycline).

B) Les annotations d'*E. coli* K-12 ne donnent aucune indication sur la fonction des CDS *ybcK*, *ybcL* et *ybcM* (*unknown*). En revanche, YbcK montre une faible similitude avec une putative recombinase, YbcL avec une protéine périplasmique et YbcM avec un régulateur transcriptionnel.

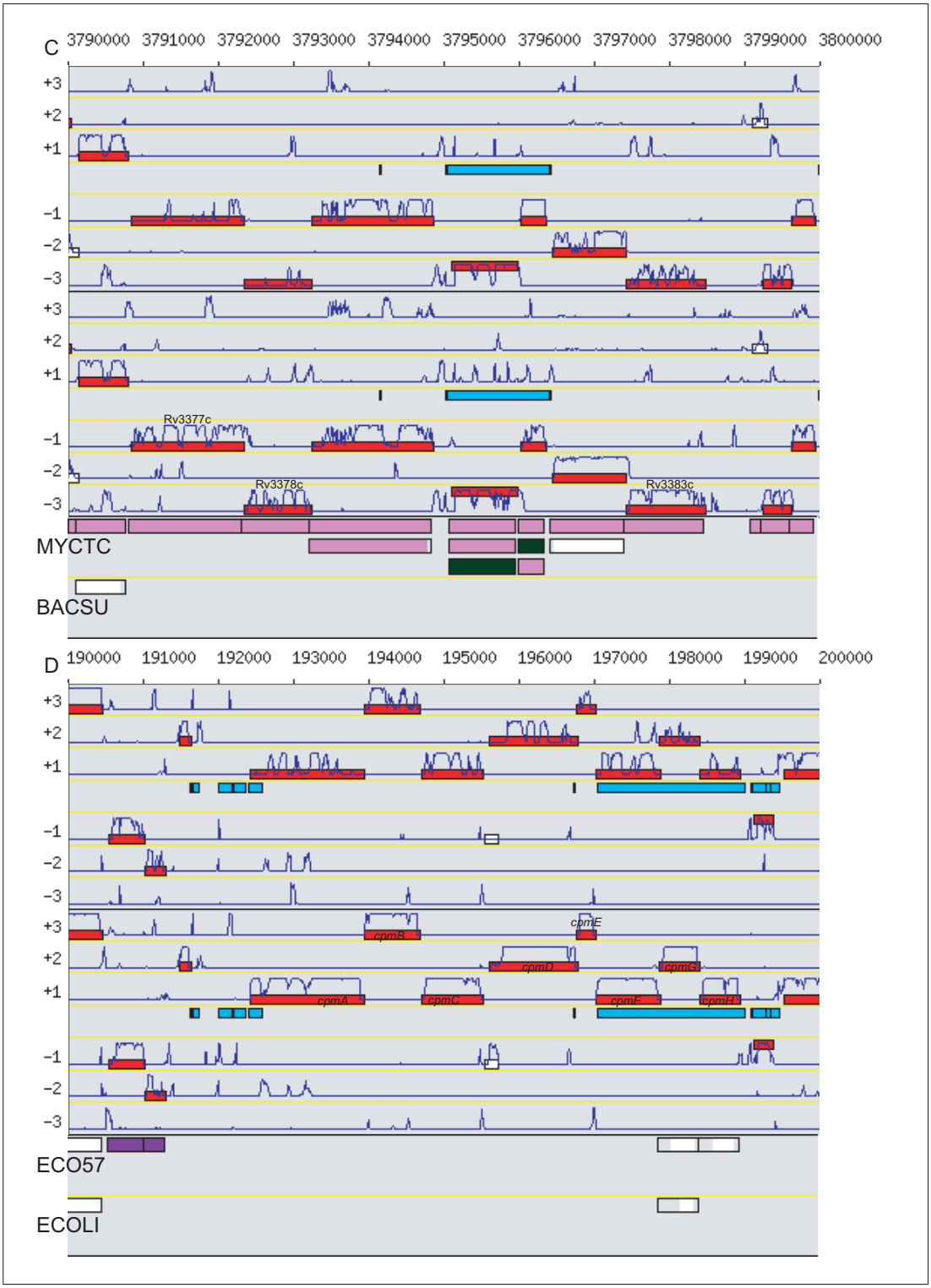


FIG. 7.8 – Validation des matrices de transition en fonction des k classes de gènes

C) Les CDS de *M. tuberculosis* H37Rv, *Rv3377c*, *Rv3378c* et *Rv3383c*, sont annotées respectivement comme cyclase putative, protéine hypothétique et polyprenyl synthétase IDSB.

D) Les protéines de *P. luminescens* codées par *cpm[ABCDEFGH]* sont impliquées dans la biosynthèse et la résistance à l'antibiotique carbapenem.

Définition des objets sur les cartes génomiques : CDS rouge, '*common*' ; CDS violette, '*uniqBank*' ; CDS rose, '*noStatusBank*' ; CDS orange, '*suspiciousBank*' ; CDS jaune, '*wrongBank*' ; CDS blanche, '*uniqAGC*' ; CDS verte, '*noStatusAGC*' ; CDS turquoise, '*ambiguousAGC*' ; CDS bleue, '*newAGC*'. Les rectangles turqoises non phasés sont de longues répétitions prédites par Nosferatu [Achaz *et al.*, 2002]. Les CDS qui sont placées en haut de la phase contiennent une mutation qui peut générer un seul fragment (CDS partielle), deux fragments phasés (codon stop en phase), un ou plusieurs fragments non phasés, décalés par rapport à la courbe de probabilité de codage (décalage du cadre de lecture). Dans chacun de ces quatre exemples, la troisième carte représente les synténies prédites entre les CDS du génome pivot et celles d'autres génomes [Labarre & Médigue, 2004]. Un rectangle blanc représente un orthologue isolé (absence de synténie). Un rectangle de couleur représente un orthologue colocalisé (présence de synthénie). La partie transparente d'un rectangle blanc ou de couleur représente la partie de la protéine sujette qui ne s'aligne pas avec la protéine requête du génome pivot.

de la figure 7.8 p. 232, les CDS des banques sont correctement prédites par *AMIGene* (elles sont toutes en rouge). La probabilité moyenne de codage de Rv3378c est de 0,64 avec la matrice III construite à partir d'une partition en quatre classes, alors qu'elle n'était que de 0,27 avec la matrice III construite à partir d'une partition en trois classes.

Enfin, l'exemple de *P. luminescens* permet de constater que les matrices construites à partir de la variante d'*AMIMat*, qui utilise uniquement la séquence chromosomique, sont de qualité équivalente aux matrices construites à partir de la variante qui utilise en plus, le jeu de CDS annotées dans les banques.

7.5 Originalités et limites actuelles de la stratégie *AMIMat*

La stratégie *AMIMat* permet d'améliorer la prédiction de gènes par chaînes de Markov puisque la qualité des matrices de transition est une condition *sine qua non* pour la précision de la phase de reconnaissance des CDS [Azad & Borodovsky, 2004]. Par exemple, les gènes de capsule de *N. meningitidis* Z2491 (Serogroup A) étaient ratés par *AMIGene* lorsque l'on utilisait une seule matrice de transition et ne le sont plus avec trois matrices (Fig. 7.1 p. 198). A l'inverse de *GeneMark-Genesis*, *AMIMat* est semi-automatique, paramétrable par l'utilisateur (il n'y a pas systématiquement trois classes), souple (intrinsèque ou extrinsèque), couramment utilisée (vingt-cinq jeux de matrices sont disponibles) et repose sur des programmes libres (*Prokov*, *AFCcodon*, *Ether*). De plus, dans le cas de *GeneMark-Genesis*, le « partionnement » automatique est effectué sur les valeurs *RSCU* des gènes, alors que dans le cas d'*AMIMat*, il est réalisé sur les coordonnées de l'AFC des *RSCU* des gènes.

7.5.1 stratégie bioinformatique

Une amélioration évidente de la stratégie *AMIMat* est qu'elle soit complètement intégrée dans la plate-forme *Genostar*. Actuellement seules les étapes RSCU–AFC–Ether sont effectuées dans le module *GenoBool*. Les autres étapes pourraient se dérouler dans le module *GenoAnnot* (Fig. 7.3 p. 214). Il serait alors exploiter l'interopérabilité des modules de cette plate-forme.

A priori, on aurait tendance à délimiter des classes de gènes à partir des fréquences relatives de codons puis à construire des matrices à partir des fréquences des trinucléotides (ordre 2). Cependant, l'heuristique que nous utilisons actuellement (classes de gènes à partir des valeurs RSCU et apprentissage d'ordre 5) donne de meilleurs résultats. Il serait donc intéressant de fournir une explication formelle sur ce point. Il existe une méthode alternative pour modéliser l'hétérogénéité de composition des CDS qui consiste à utiliser des chaînes de Markov cachées (voir p. 105 [Besemer *et al.*, 2001, Larsen & Krogh, 2003, Nicolas, 2003]).

La partie la plus délicate de la stratégie *AMIMat* est le choix du nombre de classes, de la meilleure partition et l'interprétation biologique de ces résultats. D'un point de vue méthodologique, en ce qui concerne le choix des indices de codon caractérisant les gènes, nous devrions tester différentes mesures : la fréquence absolue du codon, la fréquence relative du codon, la fréquence

relative du codon synonyme, le RSCU (voir p. 89). Actuellement, nous utilisons les deux dernières valeurs. Certains auteurs pensent qu'en utilisant le RSCU, nous perdons de l'information (*e.g.* biais de la composition en acides aminés) et que les résultats sont donc parfois moins bons [Perrière & Thioulouse, 2002]. Sachant que notre objectif est la prédiction de gènes dans les séquences nucléiques, l'utilisation des valeurs RSCU permet justement d'étudier les biais de codons synonymes dans les séquences nucléiques, sans se préoccuper des séquences protéiques ; en particulier, cette approche améliore la prédiction des gènes de composition atypique. De plus, certains biais en acides aminés transparaissent déjà au niveau nucléique (*e.g.* différencier deux protéines hydrophobes, l'une cytoplasmique l'autre membranaire ; communication personnelle S. Cruveiller). Enfin, nous envisageons d'utiliser d'autres méthodes d'analyse de données multivariées. D'une part, il existe d'autres méthodes d'analyse multifactorielle, comme l'ACP, l'analyse factorielle floue (*fuzzy* [Perrière & Thioulouse, 2002]). D'autre part, il serait nécessaire d'avoir à notre disposition d'autres méthodes de classification comme une méthode de classification hiérarchique qui fonctionnerait sur un grand nombre d'individus, une méthode de classification mixte, des nouvelles variantes de la méthode des centres mobiles, afin de s'affranchir éventuellement du nombre de classes qui doit être connu *a priori* (voir p. 133).

En particulier, A. Guenoche nous a proposé d'utiliser un K-means sélectif, progressif par densité, directement sur les données brutes afin de tenir compte de toute l'information disponible (RSCU dans le cas de l'étude de l'usage des codons synonymes des gènes). La méthode est sélective dans le sens où elle propose plusieurs façons d'évaluer la qualité des classes et de comparer les partitions [Guénoche, 2003]. Elle est progressive car à chaque étape, l'expert choisit une ou plusieurs classes, guidé par des raisons biologiques et des raisons intrinsèques (la classe choisie a des mesures de qualité meilleures que ses concurrentes [Guénoche & Lescot, 2002]). La méthode de classification par densité permet de traiter de grands tableaux de distances (n>10000) car elle ne nécessite pas de mémoriser la matrice en son entier [Guénoche, 2004].

Ainsi, en multipliant le nombre d'évidences, nous pourrions confirmer le sens biologique d'une classification de gènes en k classes suivant différents biais d'usage du code dont les principales tendances sont révélées par une analyse multifactorielle.

7.5.2 Interprétation biologique

Du point de vue de l'interprétation biologique des biais d'usage des codons qui dirigent les k classes de gènes, il serait très intéressant d'entamer des collaborations qui nous permettrait de valider expérimentalement les différentes hypothèses que nous avons émises, par exemple :

- Existe-t-il un jeu de référence de gènes dont l'origine exogène a été démontrée expérimentalement ? Sinon, comment le démontrer ? Si les gènes de classe III ont été acquis par transferts horizontaux, est ce que leur richesse en A+T à un lien avec la structure en rosace du nucléoïde et le mécanisme de transfert en lui-même ?
- Y a-t-il des gènes hautement exprimés dans la classe de gènes potentiellement acquis par transfert horizontal (III) chez *E. coli* K-12 ? Inversement, y a-t-il des gènes HGT en classe

PHX (II) chez *B. subtilis* ?

- Y a-t-il des gènes HGT G+C riches ? Si oui, alors la classe II de *P. luminescens* est-elle un mélange de gènes G+C riches, certains étant HPX, d'autres HGT, et d'autres encore à la fois HGT et PHX ?
- Les gènes essentiels de *M. tuberculosis* H37Rv et de *P. luminescens* appartiennent-ils à la classe de gènes précoces (resp. II et I) ?

Chapitre 8

Programme de prédiction de gènes bactériens : *AMIGene*

Un des travaux majeurs de cette thèse a consisté à développer une stratégie de prédiction de régions codantes qui permet soit d'annoter un génome procaryote nouvellement séquencé soit de réannoter un génome dont les annotations sont répertoriées dans les banques de séquences (*Article II* p. 274 [Vincent, 2001]). Cette stratégie d'annotation des gènes microbiens, *Annotation of MIcrobial Genes* (*AMIGene*), est divisée en deux phases : (i) la phase de reconnaissance des CDS, fondée sur les modèles de séquences d'ADN par chaînes de Markov (voir p. 92) et (ii) la phase de filtrage de ces CDS, fondée sur l'expertise des annotateurs. *AMIGene* intègre des annotations syntaxiques des CDS de la même manière que procéderait un annotateur pour valider une CDS d'après une série de décisions. Une procédure classique d'annotation manuelle consiste à parcourir une carte graphique de la séquence chromosomique synthétisant les annotations d'objets génomiques issus de diverses analyses automatiques, pour identifier visuellement la structure potentielle des gènes et des opérons (*e.g.* cohérence entre la présence d'une CDS, son potentiel de codage, ses résultats de similitude et son environnement génomique).

8.1 Objectifs de la stratégie *AMIGene*

Les membres du groupe Atelier de Génomique Comparative sont à la fois programmeurs et utilisateurs de serveurs de programmes d'analyse des séquences procaryotes (*AMIGene* pour la prédiction de CDS), d'interfaces d'annotation (*MaGe* adossé à PkGDB), de plates-formes exploratoires (*Genostar* successeur d'Imagene). *AMIGene* a été développé initialement dans Imagene. Ce système fournit des interfaces utilisateurs qui permettent de prédire des objets génomiques, de caractériser les gènes et leur produit, en exécutant différentes stratégies d'analyse automatique, puis de superposer ces résultats dans une même carte graphique du fragment de séquence chromosomique. Par exemple, les phases ouvertes de lecture (délimitées par un codon d'initiation et un codon de terminaison en phase) peuvent être superposées aux courbes de probabilités de codage

calculées par le programme *GeneMark*.

Les CDS les plus probables peuvent ainsi être mises en évidence, comme le montre l'exemple de la figure 8.1 p. 240. L'utilisateur va alors choisir les plus longues CDS qui présentent les plus fortes probabilités de codage, en cherchant à minimiser les recouvrements locaux entre CDS adjacentes et à maximiser la couverture globale du chromosome. Cette sélection n'est pas toujours aisée, notamment chez les génomes G+C riches, dans les régions atypiques (relativement A+T riches par rapport au reste du génome), dans des régions de décalage du cadre de lecture des CDS, dans des régions de répétitions, ou encore dans le cas où il est nécessaire de réajuster le codon d'initiation. L'utilisateur doit aussi se méfier des pics de codage parasites (montée artefactuelle de la courbe de probabilités).

Dans l'exemple du fragment de *M. tuberculosis* H37Rv (Fig. 8.1 p. 240), le premier problème concerne le choix d'annoter la CDS1 et la CDS2, ou bien d'annoter Rv1946c (et dans ce cas, la position de son codon d'initiation doit être réajustée afin de minimiser son chevauchement avec Rv1947). La CDS1 en phase +3 et la CDS2 en phase +1 sont incluses dans Rv1946c en phase -1. Jusqu'à preuve du contraire, le cas d'inclusion d'une CDS dans une autre, lorsqu'elles sont en sens contraire, n'existe pas dans l'ADN « strictement » bactérien, tandis que cela aurait été observé chez les bactériophages[1]. Comme l'ADN des phages peut être intégré de manière stable dans l'ADN bactérien, il devrait être possible d'observer un cas d'inclusion de CDS en sens contraire, au niveau de régions phagiques d'un chromosome bactérien. En revanche, le cas d'inclusion de CDS dans la même orientation peut être observé dans l'ADN « strictement » bactérien (cela reste un événement rare). Il s'agit soit de deux CDS complètes comme dans le cas de l'inclusion de *comS* dans *sfrAB* chez *B. subtilis* [Hamoen *et al.*, 1995], soit de deux fragments de la même CDS créés par un décalage du cadre de lecture compensé (double mutation authentique ou erreur de séquençage [Rojas *et al.*, 2003]). Dans l'exemple de la figure 8.1 p. 240, Rv1946c présente une probabilité moyenne de codage (Pc) plus faible que celles de la CDS1 et de la CDS2 : il semble donc raisonnable d'éliminer Rv1946c et de sélectionner la CDS1 et la CDS2 (ce qui doit être confirmé par les résultats d'autres analyses comme les recherches de similitude dans des banques de séquences et de motifs protéiques).

Le deuxième problème concerne le choix d'annoter Rv1947 ou la CDS3 (cas d'inclusion de CDS sens contraire). Vu la probabilité moyenne de codage (Pc) et la longueur (L) de la CDS3 par rapport à celles de Rv1947, il semble raisonnable de supprimer la CDS3 et de conserver Rv1947. Si la Pc de la CDS3 était plus importante, alors on l'aurait aussi conservée car sa longueur est non négligeable par rapport à celle de Rv1947, et ce cas d'inclusion aurait pu pointer sur une caractéristique importante de la séquence comme une région phagique ou une CDS fantôme. Les résultats d'analyses ultérieures permettraient alors de trancher.

Le troisième problème concerne la présence d'un trou d'annotation généré par un décalage du cadre de lecture du type présence d'une prédiction de codage en absence de CDS (ellipse verte de la figure 8.1 p. 240). Ce type de *frameshift* peut être révélateur d'une erreur de séquençage ; ce qui

[1] Je n'ai cependant pu trouvé aucun exemple le confirmant

est le cas ici puisque ce décalage du cadre de lecture (et donc ce trou d'annotation) a disparu de la dernière version du chromosome mais la nouvelle CDS, qui se termine à la position 2198627, n'a pas été annotée [Camus *et al.*, 2002].

Le quatrième problème est la faible prédiction de codage des gènes Rv1948c, Rv1949c, Rv1950c, Rv1951c et Rv1952 (notamment, Rv1949c en phase -3 a une probabilité de codage qui oscille) ; ceci est généralement révélateur d'une matrice de transition de mauvaise qualité (effectivement, avec les quatres matrices de transition spécifiques des classes d'usage des codons synonymes du génome de *M. tuberculosis* H37Rv, les prédictions sont meilleures ; voir p. 220).

Enfin, le dernier problème concerne le choix d'annoter la CDS4 et la CDS5 ou bien d'annoter Rv1954c (cas d'inclusion de CDS sens contraire). Bien que Rv1954c soit la plus longue des trois CDS, c'est aussi elle qui présente la plus faible Pc ; elle est donc éliminée. Sans information supplémentaire, on est incapable de savoir si l'inclusion de la CDS5 dans la CDS4 est révélatrice d'un décalage du cadre de lecture compensé, s'il est nécessaire de réajuster la position du codon d'initiation de la CDS4 afin d'éliminer l'inclusion ou si, tout simplement, la CDS5 n'existe pas (pic parasite).

Ainsi, dans les régions où l'annotation syntaxique de CDS est délicate, une méthode intrinsèque automatique doit pouvoir conserver les CDS sur lesquelles portent l'ambiguïté, laissant le soin à l'expertise humaine de trancher manuellement à l'aide d'annotations fonctionnelles issues d'analyses ultérieures. Dans ce cadre d'un projet d'annotation d'un nouveau génome, l'annotation automatique sert en fait de *préannotation* et nécessite d'être suivie par plusieurs passages d'annotation manuelle. Il nous paraît donc plus grave que la méthode automatique oublie une « vraie » CDS (*i.e.* faux-négatif), plutôt qu'elle prédise une « fausse » CDS (*i.e.* faux-positif). Dans l'exemple de la figure 8.1 p. 240, nous souhaitons sélectionner automatiquement les CDS suivantes : CDS1, CDS2, Rv1947, Rv1948c, Rv1949c, Rv1950c, Rv1951c, Rv1952, Rv1953, CDS4, CDS5 et Rv1955. Un annotateur expert pourra alors décider *a posteriori*, à l'aide de résultats d'autres méthodes comme des recherches de similitude dans des banques protéiques, s'il conserve ou non les CDS Rv1948c et Rv1951c, et s'il réajuste ou non la position du codon d'initiation de la CDS4 et de Rv1955.

Partant de ces observations et des quelques limites des programmes *GLIMMER* et *GeneMark* mentionnées dans la section précédente, nous avons développé une stratégie de sélection automatique de CDS sur une séquence génomique procaryote dont les *objectifs principaux sont les suivants* :

1. Tenir compte de l'hétérogénéité de composition en oligonucléotides des séquences génomiques analysées, dès la phase d'apprentissage des CDS, en construisant une matrice de transition pour chaque classe de CDS caractérisée en fonction de l'usage des codons synonymes (au lieu d'une seule matrice de transition par génome ; voir p. 197).

2. Réajuster si nécessaire la position du codon d'initiation des CDS (*e.g.* au moment du calcul de la Pc).

3. Parmi l'ensemble des CDS précédemment définies, sélectionner les plus probables (*i.e.* les CDS plus longues qui possèdent les Pc les plus élevées en minimisant les recouvrements entre CDS adjacentes et en maximisant la couverture du génome).

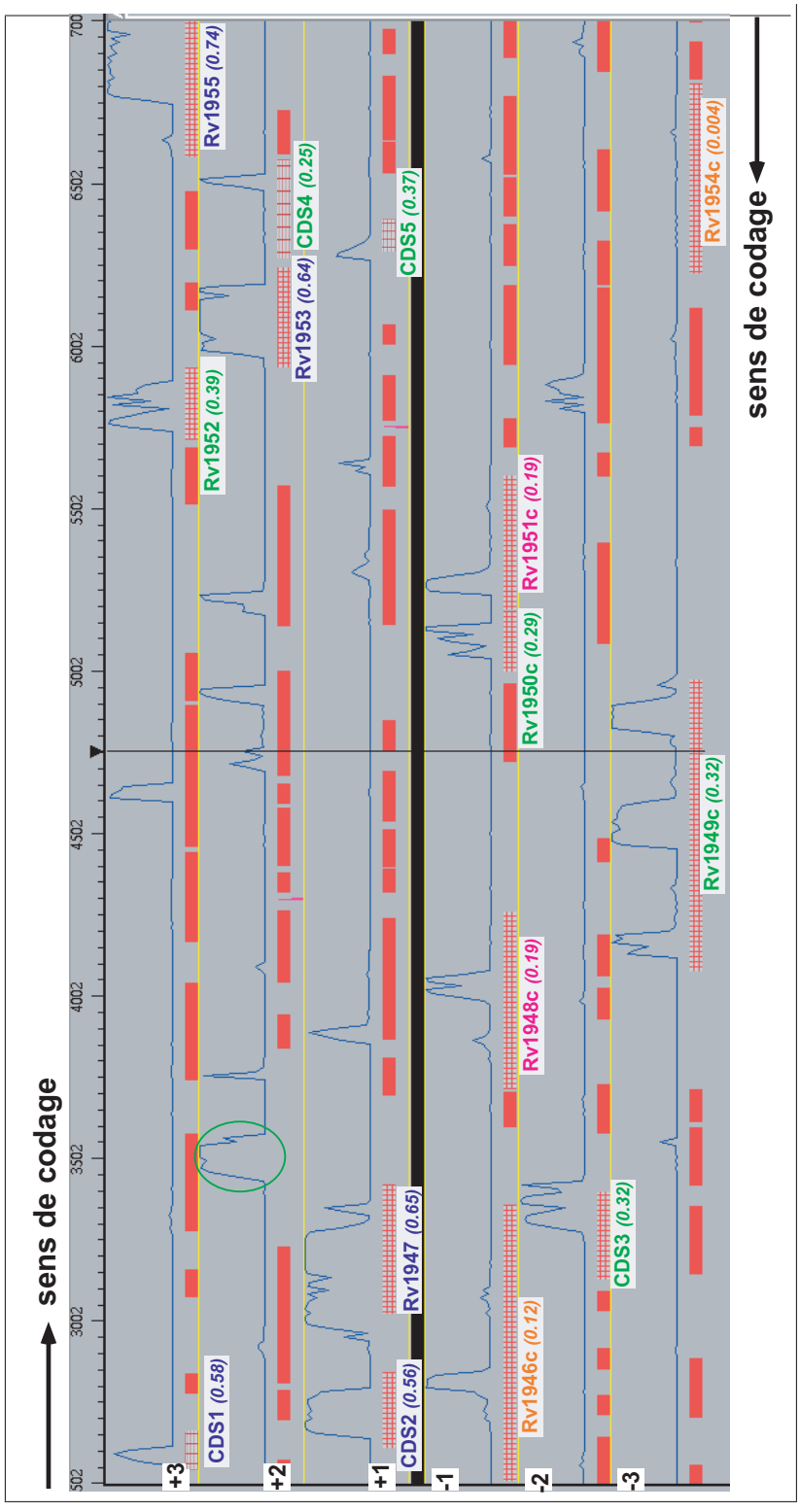


FIG. 8.1 – Annotation manuelle de CDS à partir d'une représentation cartographique synthétisant les résultats de différentes analyses

Cette représentation cartographique a été obtenue avec la plate-forme Imagene [Médigue *et al.*, 1999a] après analyse d'un fragment du chromosome de *M. tuberculosis* H37Rv. Les rectangles rouges représentent les CDS prédites par une recherche par signal (*SPOC* ; voir p. 268). Les courbes bleues représentent la probabilité de codage calculée par une recherche par contenu (*GeneMark*). Les résultats ont été superposés sur chacune des six phases de la séquence nucléique. Les CDS sélectionnées sont représentées par des rectangles quadrillés rouges. Ce sont les plus longues CDS qui ont les plus fortes Pc et qui couvrent presque complètement le fragment chromosomique (recouvrements minimum entre les CDS). L'ellipse verte met en évidence un décalage du cadre de lecture. Le label des CDS est suivi de la *(Pc)* : orange ($Pc < 0,15$), rose ($0,15 \leq Pc < 0,25$), vert ($0,25 \leq Pc < 0,50$) et bleue ($Pcgeq0,5$). Les CDS dont le label commence par Rv sont celles annotées dans GenBank. On observe deux conflits d'annotation : sachant qu'il est possible de réajuster le codon d'initiation, doit-on annoter CDS1, CDS2 et Rv1947 ou Rv1946c et CDS3 ? doit-on annoter CDS4 et CDS5 ou Rv1954c ? Pour trancher, il est nécessaire d'ajouter les résultats d'autres analyses, comme la recherche de similitude dans les banques de séquences protéiques.

4. Filtrer ces CDS en plusieurs étapes, l'analyse devenant de plus en plus sévère sur des CDS de moins en moins probables.
5. La stratégie doit être paramétrable de façon à pouvoir éviter les faux-négatifs (paramètres peu sévères), ou bien les faux-positifs (paramètres très sévères), ou à trouver une situation intermédiaire en fonction de ce que l'utilisateur désire (analyse de la séquence minutieuse ou grossière).

8.2 Etapes de la stratégie *AMIGene*

La stratégie *AMIGene* se déroule en deux phases : (i) reconnaissance de CDS caractérisées par leurs Pc en fonction des différentes matrices de transition du génome étudié et (ii) heuristique de sélection des CDS les plus probables que nous décrivons ici en détail. Ces deux étapes permettent d'enchaîner différents modules de reconnaissance et de post-traitement des CDS (voir p. 207 et Fig. 8.2 p. 242).

8.2.1 Reconnaissance de CDS

Le programme *AMIGene* utilise des modules présentés dans le chapitre précédent et les modules suivant :

1. *SPat* est un programme de recherche d'expressions régulières dans les séquences biologiques (*Search Pattern* vient de la plate-forme Imagene ; voir p. 80). Dans le cadre de nos analyses, nous l'utilisons pour rechercher la position de tous les codons d'initiation sur la séquence d'ADN ([ACGT]TG).
2. *compute_Pc* combine les résultats de *prokov_orf* (liste des *Leftmost Start* LS_CDS de longueur supérieure à 60 pb dans les six phases de lecture de la séquence d'ADN), de *SPat* (liste des codons d'initiation de la séquence) et de *prokov_curve* (six vecteurs de probabilités de codage calculées à partir d'une matrice de transition et de la séquence) pour calculer la probabilité moyenne de codage (Pc) des CDS, et réajuster la position du codon d'initiation. Le choix du codon le plus en 5′ appliqué par *prokov_orf* peut générer de trop longues CDS, ce qui a pour conséquence de diminuer la Pc et d'augmenter la taille des recouvrements avec les CDS adjacentes. C'est pourquoi *compute_Pc* met en œuvre une heuristique de réajustement de la position du codon d'initiation des LS_CDS (Fig. 8.3 A p. 244). D'abord, nous cherchons en 3′ de la position du LS la première position de la séquence pour laquelle la probabilité de codage est supérieure au seuil *climb_P* (*e.g.* 0,965). Ensuite, nous cherchons en 5′ de cette position, la position du premier codon d'initiation en phase dans la liste des codons d'initiation fournie par *SPat* : c'est un codon d'initiation alternatif appelé *AMIGene Start* (AS). Enfin, nous calculons la Pc de la LS_CDS (LS_Pc) et la Pc de la AS_CDS (AS_Pc). Si la différence entre ces deux Pc est supérieure à un certain seuil (*e.g.* AS_Pc - LS_Pc >= *diff_Pc*), alors on conserve les deux positions de codon d'initiation et les deux Pc (sinon AS = LS). Pour

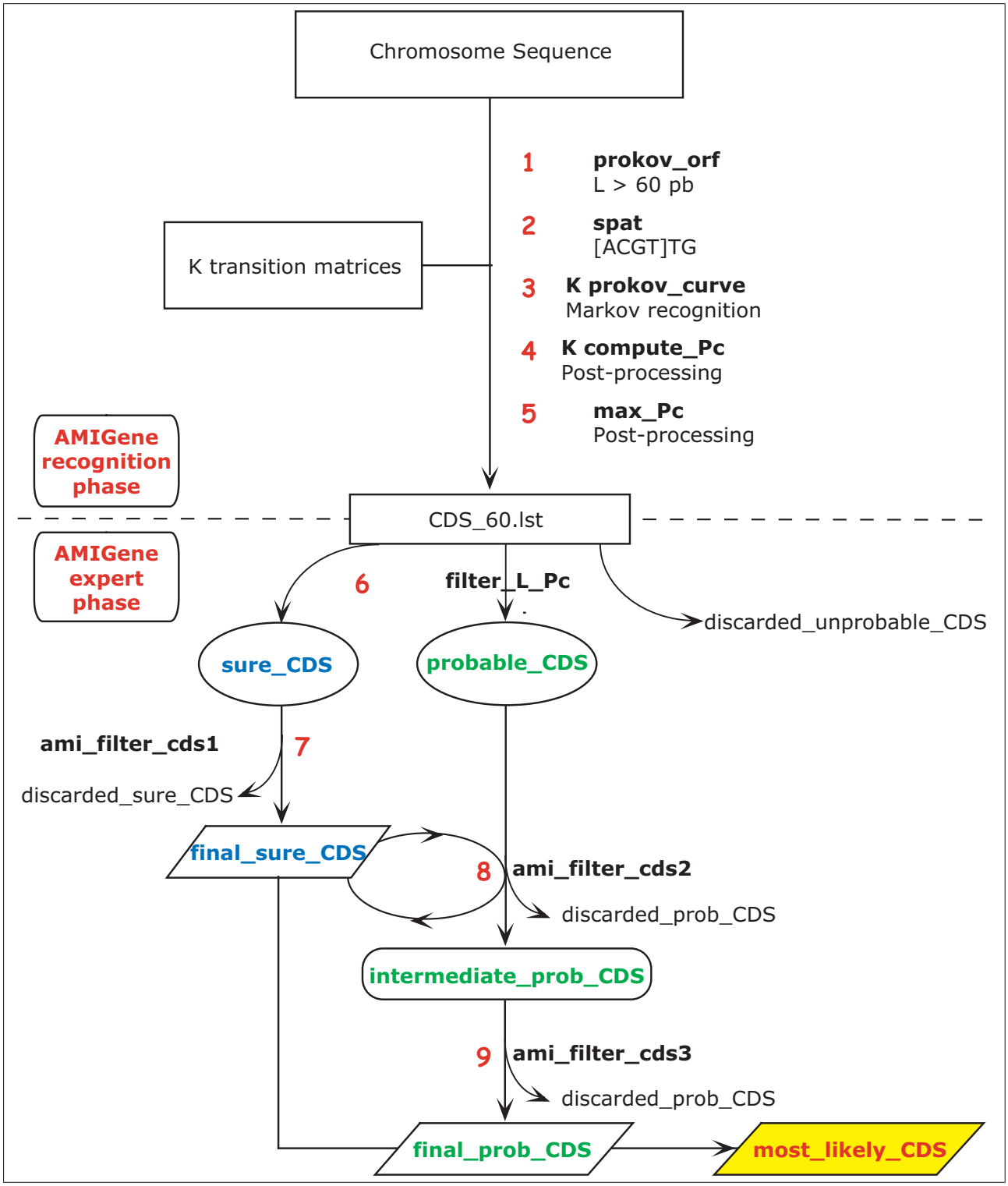


FIG. 8.2 – Stratégie *AMIGene*

Stratégie de prédiction des CDS les plus '*probable*' d'un génome procaryote. La première série d'étapes permettent de définir un jeu de CDS initiales caractérisées par leurs positions de début et de fin, leur longueur, leur meilleure probabilité moyenne de codage (Pc) obtenue avec les k matrices. La seconde série d'étapes consiste à filtrer les CDS de la même manière qu'un annotateur valide manuellement les CDS selon son expertise. On sépare les CDS en trois listes : la liste des CDS '*sure*' dans lesquelles on peut avoir confiance, la liste des CDS '*probable*' qui révèlent des situations conflictuelles et la liste des CDS '*improbable*' que l'on élimine dès le départ. Les étapes 6, 7 et 8 sont décrites plus précisément dans les figures 8.4 p. 245 et 8.5 p. 248.

que le réajustement de la position du codon d'initiation de la CDS soit effectivement pris en compte, une contrainte supplémentaire est ajoutée : le réajustement ne doit pas entraîner l'élimination ultérieure de la CDS (Fig. 8.3 p. 244 et voir p. 259). Par exemple, si AS_L est inférieure à 63 pb, on ne réajuste pas le codon d'initiation.

3. *max_Pc* permet de faire une synthèse des résultats ; en effet, pour chaque CDS, il conserve la meilleure AS_Pc, la LS_Pc et le numéro de matrice correspondant. On fait ici l'hypothèse que la matrice de transition la mieux adaptée à l'usage des codons synonymes d'une CDS permettra le calcul de sa Pc la plus élevée.

Afin de définir un jeu initial de CDS, les modules *prokov_orf*, *SPat*, *prokov_curve*, *compute_Pc* et *max_Pc* sont exécutés sur la séquence génomique étudiée (étapes 1 à 5 de la figure 8.2 p. 242). Si k matrices de transition sont disponibles, *prokov_curve* et *compute_Pc* seront exécutés k fois. Ainsi chaque CDS est définie par son identifiant, sa position de début sur la séquence d'ADN, sa position de fin et son orientation (directe ou inverse), et caractérisée par sa phase de lecture, la position du codon d'initiation alternatif proposé par *AMIGene*, AS_L, LS_L, AS_Pc, LS_Pc et le numéro de la matrice de transition correspondant à AS_Pc (la plus élevée).

8.2.2 Filtrage des CDS

Dans une seconde phase, qui permet de filtrer le jeu de CDS précédemment défini, *AMIGene* enchaîne les modules *filter_L_Pc*, *AMI_filter_CDS*1, *AMI_filter_CDS*2, *AMI_filter_CDS*3 (étapes 6 à 9 de la figure 8.2 p. 242).

D'abord, *filter_L_Pc* permet de regrouper les CDS en deux listes (Fig. 8.3 B p. 244 et voir p. 250) :

1. La liste des CDS '*sure*' regroupe les CDS dont la AS_Pc est supérieure à *sure_Pc* et dont la AS_L est supérieure à *sure_L*.
2. La liste des CDS '*probable*' regroupe les CDS dont la Pc est comprise entre *prob_Pc* et *sure_Pc* et dont la AS_L est supérieure à *prob_L*.

Au passage, les autres CDS, de statut '*improbable*', sont éliminées (Fig. 8.3 B p. 244). Ce premier filtre permet d'éliminer simplement un certain nombre de faux-positifs, *i.e.* le bruit généré par les CDS de n'importe quelle longueur dont la Pc est faible, les petites CDS dont la Pc est moyenne, et les très petites CDS dont la Pc est élevée (Fig. 8.3 B p. 244 ; ces dernières n'avaient de toute manière pas été définies par *prokov_orf* avec L > 60 pb).

Puis l'heuristique *AMI_filter_CDS* se déroule en trois étapes de filtrage, qui combinent des critères de plus en plus sévères sur des CDS de moins en moins probables (Fig. 8.4 p. 245). Cette heuristique vise à éliminer des faux-positifs suivant des critères de longueur, de probabilité moyenne de codage et de recouvrements entre CDS adjacentes.

L'étape *AMI_filter_CDS*1 se préoccupe des cas d'inclusion dans la liste des CDS '*sure*' (étape 7 de la figure 8.4 p. 245 et Fig. 8.5 A p. 248). Pour chaque CDS incluse, le pourcentage d'inclusion est calculé (*i.e.* le rapport des longueurs de la plus petite CDS sur la plus longue). Si ce pourcentage

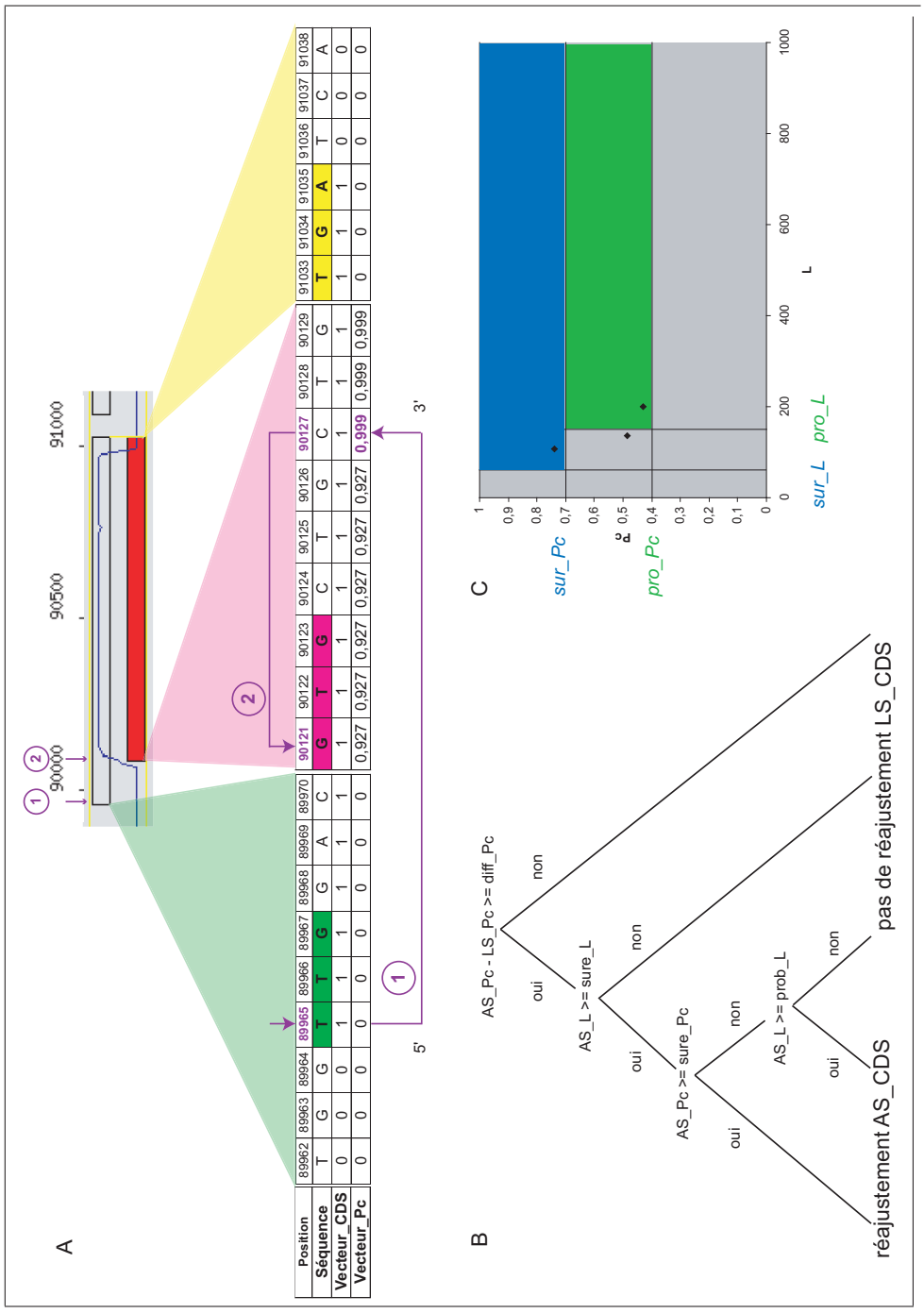


FIG. 8.3 – Réajustement du *start* et filtre des CDS sur la longueur et la Pc

A) Réajustement automatique de la position du codon d'initiation de la CDS *mraW* d'*E. coli* K-12. Premièrement, la LS_CDS est représentée par le rectangle transparent, elle commence par le codon d'initiation TTG à la position 89965. La première position de la séquence pour laquelle la probabilité de codage est supérieure au seuil de 0,965 (*climb_P*) correspond à la position 90127. Deuxièmement, nous cherchons en 5′ de la position 90127 la position du premier codon d'initiation en phase : c'est le codon d'initiation GTG à la position 90121 de la AS_CDS (rectangle rouge). Dans cet exemple, nous considérons uniquement les codons d'initiation ATG, GTG et TTG (et pas CTG). Le codon d'initiation a été trop réajusté, sachant que le codon d'initiation annoté dans EcoGene17 est l'ATG à la position 90094.

B) Le réajustement de la position du codon d'initiation de la CDS ne doit pas entraîner son élimination ultérieure.

C) Les seuils longueur de CDS (*e.g.sure_L* = 60 pb et *prob_L* = 150) et de Pc (*e.g.sure_Pc* = 0,7 et *prob_Pc* = 0,4) permettent de définir trois statuts de reconnaissance : les CDS '*improbable*', '*probable*' et '*sure*' sont regroupées, respectivement dans les surfaces grise, verte et bleue. Imaginons par exemple une LS_CDS avec une longueur de 201 pb et une Pc de 0,42 (point noir de la surface verte). Elle est '*probable*' et sera donc conservée par *filter_L_Pc*. Si la AS_CDS correspondant à cette LS_CDS a une AS_L de 139 et une AS_Pc de 0,49, alors elle sera '*improbable*' et *filter_L_Pc* l'éliminera à cause de la longueur. Dans ce cas, il ne faut pas réajuster la position du codon d'initiation de la CDS (point noir de la surface grise). En revanche, si la AS_CDS correspondante a une AS_L de 102 et une AS_Pc de 0,73 alors elle sera '*sure*' et *filter_L_Pc* la conservera. Dans ce cas, il faut réajuster le codon d'initiation (point noir de la surface bleue).

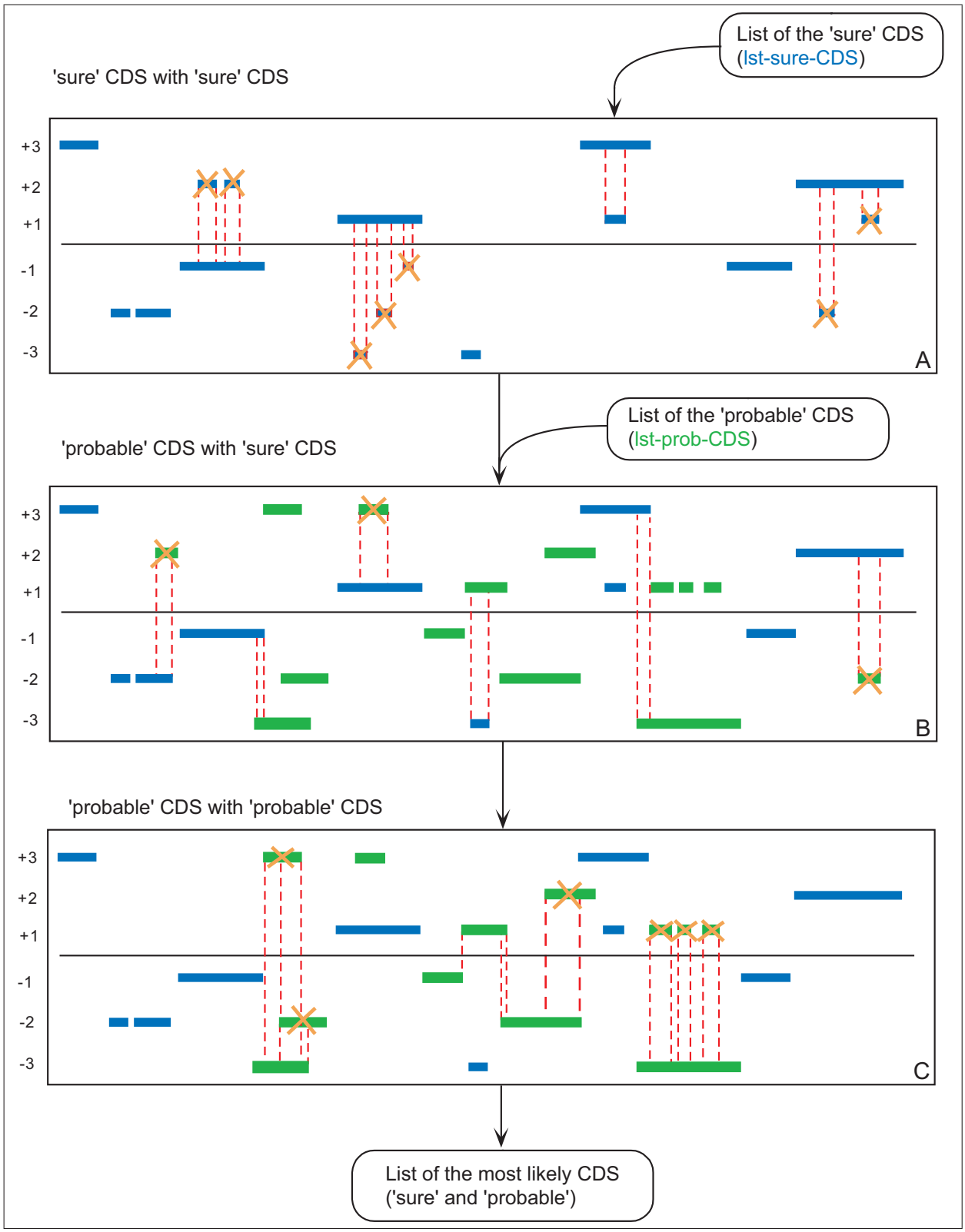


FIG. 8.4 – Etape de post-traitement : l'heuristique d'*AMIGene*

Les CDS ont été séparées en trois groupes : '*improbable*', '*probable*', '*sure*'. La première étape analyse les cas d'inclusion entre CDS '*sure*'. La deuxième étape analyse les cas de recouvrement entre CDS '*sure*' et '*probable*'. La troisième étape analyse les cas de recouvrement entre CDS '*probable*'.

est supérieur à *sure_I* (voir p. 250), c'est-à-dire si la taille de la petite CDS est non négligeable par rapport à celle de la longue, la petite CDS (incluse) est conservée (la longue CDS est bien entendu aussi conservée). La valeur du paramètre seuil *sure_I* est plus sévère lorsque les deux CDS sont sur les brins opposés (*sure_os_I*) que lorsqu'elles sont sur le même brin (*sure_ss_I*), pour les raisons évoquées dans la section précédente.

L'étape *AMI_filter_CDS*2 (étape 8 de la figure 8.4 p. 245 et Fig. 8.5 B p. 248) repère les recouvrements entre une CDS '*sure*' (provenant de la liste issue de la sous-étape précédente) et une CDS '*probable*' qui est, le cas échéant, éliminée. Dans le cas d'une CDS '*sure*' incluse dans une CDS '*probable*', la '*probable*' est conservée. Dans le cas d'une CDS '*probable*' incluse dans une CDS '*sure*', la '*probable*' est éliminée. Dans le cas d'un chevauchement, le pourcentage de chevauchement est calculé (*i.e.* le rapport de longueurs du chevauchement sur la CDS '*probable*'). S'il est supérieur à *sure_prob_O*, la CDS '*probable*' est éliminée, sinon elle est conservée (voir p. 250).

L'étape *AMI_filter_CDS*3 (étape 9 de la figure 8.4 p. 245 et Fig. 8.5 C et D p. 248) permet d'éliminer des faux-positifs dans des régions où il n'y a que des CDS '*probable*'. Elle est basée sur le calcul du score de recouvrement *global* de chaque CDS '*probable*'. Dans cette dernière étape, il est important de ne pas éliminer les CDS à mesure que l'on avance sur le génome : le calcul d'un score de recouvrement global impose en effet de ne jamais perdre d'information sur les recouvrements existants (donc de ne pas éliminer de CDS dans un premier temps). C'est pourquoi on parcourt une première fois la liste des CDS '*probable*', pour attribuer à chacune un score total de recouvrement qui n'est autre que la somme de chacun de ses pourcentages de recouvrement. Puis, on parcourt une seconde fois la liste pour éliminer toutes les CDS '*probable*' dont le score est supérieur à *prob_glob_IO* (Inclusion–*Overlap* (IO) ; voir p. 250).

Ainsi, à l'issu de ces trois étapes de filtrage des CDS prédites, la méthode *AMIGene* fournit une liste de CDS les plus probables qui couvre au maximum le génome étudié en minimisant les recouvrements entre CDS.

8.3 Optimisation experte de la valeur des paramètres

Nous avons vu que la phase d'apprentissage est cruciale pour la phase de reconnaissance de notre stratégie de prédiction de CDS. Tout aussi crucial, c'est le choix de la valeur des paramètres. L'avantage de notre méthode de prédiction de gènes *AMIGene*est que les paramètres sont interprétables par l'utilisateur (ils ont un sens biologique). Les paramètres d'*AMIGene*, ajustés correctement, permettent de prédire un jeu de CDS les plus probables pour tout type de génome procaryote (Table 1 de l'*Article II* p. 274 et Fig. 9.1 p. 277). La valeur de ces paramètres a été déterminée dans un premier temps de façon empirique[2], puis nous avons procédé, dans un second temps, à une validation automatique de ces valeurs par rapport à un jeu de référence.

Le paramétrage peut être plus ou moins sévère en fonction de l'objectif recherché lors de l'analyse de la séquence nucléique. En d'autres termes, un paramétrage sévère prédira moins de CDS qu'un

[2]Un expert a choisi les valeurs des paramètres puis les a ajustées par essais–erreurs.

paramétrage peu sévère. Un paramétrage sévère limitera donc le nombre de faux-positifs (une CDS prédite qui n'est pas une « vraie » CDS) au détriment des faux-négatifs (une « vraie » CDS qui n'est pas prédite) et vice versa. Au sens bioinformatique, une « vraie » CDS est une CDS annotée dans un jeu de référence.

Le paramétrage dépend des caractéristiques que doit posséder le jeu de CDS que l'on cherche à définir sur une séquence d'ADN :

- Dans le cadre de nos projets de réannotation des génomes procaryotes répertoriés dans les banques de séquences nucléiques, nous avons cherché un jeu de paramètres qui puisse être appliqué à tous les génomes (afin de pouvoir comparer les résultats de réannotation) sans pour autant prédire trop de CDS (afin de repérer les erreurs d'annotation évidentes). Nous sommes donc partis d'un paramétrage sévère que nous avons ajusté empiriquement (*i.e.* aller-retour entre l'ajustement de la valeur des seuils d'*AMIGene* et l'analyse experte des résultats produits) pour le rendre moins sévère afin qu'il soit applicable à tous les génomes de notre étude (voir p. 289).
- Lors de la phase de prédiction automatique de gènes d'un projet d'annotation d'un génome nouvellement séquencé, nous voulons une annotation fine, adaptée au génome. Nous testons donc plusieurs jeux de paramètres puis une analyse experte des résultats permet de choisir un jeu de valeurs seuils. Nous cherchons la valeur des paramètres qui donnent *un* minimum de faux-positifs pour *le* minimum de faux-négatifs. En effet, les faux-positifs seront retirés au cours de la phase ultérieure de validation manuelle.

A l'issu de ces diverses optimisations expertes, nous proposons un jeu de valeurs optimales expertes de la phase de reconnaissance de CDS. Plus précisément, ces paramètres permettent de réajuster, si nécessaire, la position du codon d'initiation (*AMIGene Start* (AS)) par rapport à la position du codon le plus en $5'$ (*Leftmost Start* (LS)) :

- Le seuil de probabilité de codage, *climb_P*, est égal à 0,960. Il permet de repérer la position en 3' du LS d'une CDS correspondant à la montée de la courbe de probabilités.
- Le seuil de différence de probabilités moyennes de codage, *diff_Pc*, est égal à 0,07. Il permet d'évaluer l'impact du réajustement de la position du *start* entre la LS_CDS et la AS_CDS.
- Le seuil de probabilité moyenne de codage d'une AS_CDS '*sure*' *sure_Pc*1, est égal à 0,8. Il empêche que la AS_CDS soit éliminée ultérieurement (phase de filtrage) du fait d'une Pc trop faible.
- Le seuil de longueur d'une AS_CDS '*probable*', *prob_L*1, est égal à 150. Il empêche que la AS_CDS soit éliminée ultérieurement du fait d'une taille trop petite.

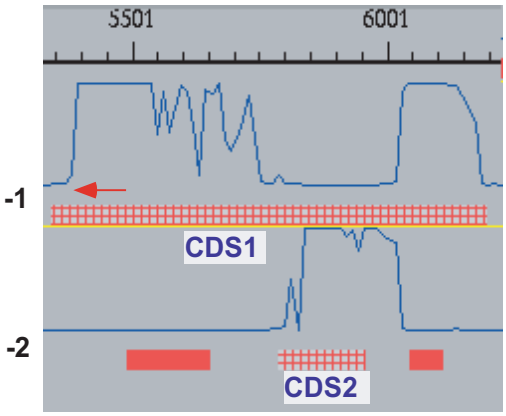
Un jeu de valeurs optimales expertes de la phase de filtrage de CDS est proposé sur le site Web d'*AMIGene* (voir p. 267) :

- Une CDS a le statut de reconnaissance '*sure*' (elle a un sens biologique) si sa longueur et sa Pc sont respectivement supérieurs à *sure_L* et *sure_Pc*2 (*e.g.* $AS\_L > 60$ pb et $AS\_Pc \geq 0,6$). Ces seuils permettent de sélectionner définitivement les CDS sans tenir compte de leur environnement avec les autres CDS (mis à part les problèmes d'inclusion).

A

Organisme	CDS	Début	Brin	Longueur	Pc	% Inclusion	Résultat
E. coli	CDS1	5337	R	855	0,51		
	CDS2	5780	R	174	0,77	20	Conservée

%inclusion
par rapport à la longueur de la plus grande CDS.
SR-inclu-sur-sens= 10%
SR-inclu-sur-sens<> 30%



B

Organisme	CDS	Début	Brin	Longueur	Pc	% Chevauchement	Résultat
E. coli	CDS1	2795	D	1536	0,58		
	CDS2	4270	R	495	0,3	CDS1 / CDS2 12	Conservée
	CDS3	4330	R	444	0,58	CDS3 / CDS2 88	Eliminée

%chevauchement
par rapport à la longueur de la CDS probable.
SR-chev-pro-sur 50%
quelque soit le sens (cas de frameshift ou de codon d'initiation trop en 5')

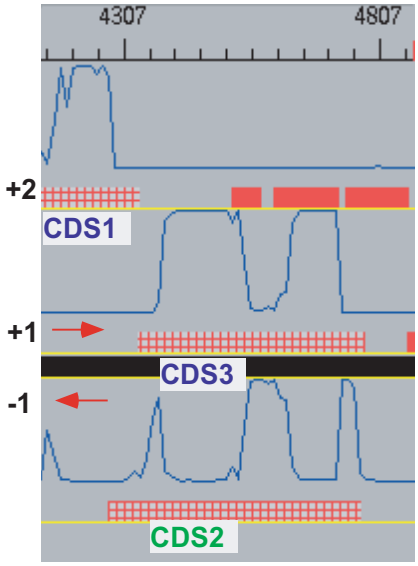


FIG. 8.5 – Exemples de recouvrements entre CDS '*sure*' et '*probable*'

A) Cas d'inclusion entre CDS '*sure*'. Les CDS1 et CDS2 '*sure*' sont dans le même sens. Le pourcentage d'inclusion de la CDS2 par rapport à la CDS1 est de 20%. Cette valeur est supérieure à *sure_ss_I* de 10%, la CDS2 est donc conservée. Si les deux CDS avaient été sur les deux brins, la CDS2 aurait été éliminée car son pourcentage d'inclusion de 20% aurait été cette fois-ci inférieure à *sure_os_I* de 30%.

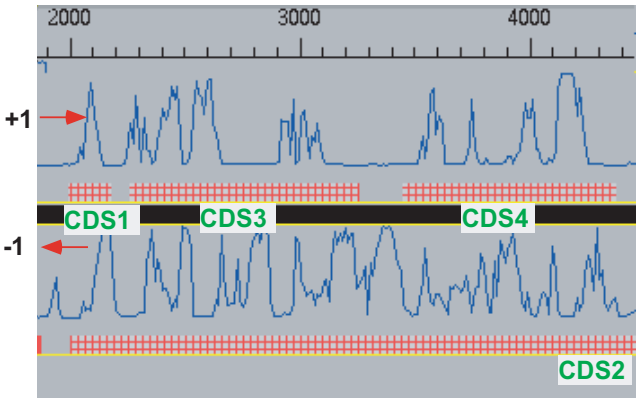
B) Cas de recouvrements entre CDS '*sure*' et '*probable*'. La CDS1 et la CDS3 sont '*sure*', la CDS2 est '*probable*' ; le chevauchement de la CDS1 avec la CDS2 est de 12%, alors qu'il est de 88% avec la CDS3. Si le seuil *sure_prob_O* est fixé à 50%, la CDS2 sera éliminée pour son chevauchement avec la CDS3.

La longueur des CDS est donnée en paires de bases.

C

Organisme	CDS	Début	Brin	Longueur	Pc	% Recouvrement	Total	Résultat
M. tuberculosis	CDS1	1991	D	186	0,21	CDS2 / CDS1 96 CDS1 / CDS2 4	96	Eliminée
	CDS2	1997	R	3918	0,37	CDS3 / CDS2 25 CDS4 / CDS2 24	53	Conservée
	CDS3	2255	D	996	0,23	CDS2 / CDS3 100	100	Eliminée
	CDS4	3437	D	933	0,23	CDS2 / CDS4 100	100	Eliminée

Score de recouvrement
somme des %chevauchement
et/ou %inclusion
avec les autres CDS_prob qui
recouvrent la CDS_courante
SR-tot-pro
80%



D

Organisme	CDS	Début	Brin	Longueur	Pc	% Recouvrement	Total	Résultat
E. coli	CDS1	4146	D	234	0,31	CDS2 / CDS1 79 CDS3 / CDS1 7	86	Eliminée
	CDS2	4194	R	228	0,35	CDS1 / CDS2 81 CDS3 / CDS2 26	107	Conservée
	CDS3	4363	R	183	0,26	CDS1 / CDS3 9 CDS2 / CDS3 32	41	Conservée

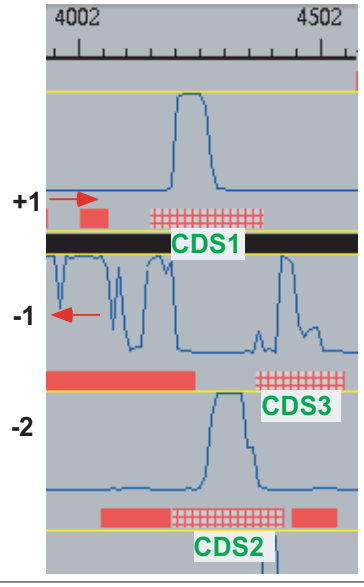


FIG. 8.5 – Exemples de recouvrements entre CDS '*probable*'

C) La CDS1 ne chevauche que la CDS2, dont seule une partie n'est représentée ; son score total de recouvrement correspond donc à son pourcentage de chevauchement avec la CDS2 (96%). Au contraire, la CDS2 possède trois recouvrements : un chevauchement avec la CDS1 qui représente 4% de la longueur de la CDS2 et des inclusions avec la CDS3 et la CDS4 qui représentent 25 et 24 % respectivement. Le recouvrement total est égal à 53%. Le score total de recouvrement de la CDS3 est de 100 puisqu'elle est incluse dans la CDS2. Si le seuil *prob_glob_IO* est de 80 les CDS1, CDS3 et CDS4 seront éliminées alors que la CDS2 sera conservée.

D) Si deux CDS ou plus vont être éliminées alors qu'elles se chevauchent, on garde celle qui a la meilleure Pc. Chacune des trois CDS possèdent des recouvrements avec les deux autres CDS. La CDS1 chevauche la CDS2 de 79.5% et la CDS3 de 7%, ce qui lui fait un score de 87%. Les CDS1 et CDS2 ont des scores supérieurs à *prob_glob_IO* mais comme elles se chevauchent et qu'on veut éviter les trous d'annotation, on n'en élimine qu'une : celle qui a la plus faible Pc. Ainsi, la CDS2 et la CDS3 seront conservées et la CDS1 sera éliminée.

- Une CDS est '*probable*' si sa longueur est supérieure à *prob_L2* et si sa Pc est comprise entre *prob_Pc* et *sure_Pc*2 (*e.g.* $AS\_L > 120$ pb et $0,3 \leq AS\_Pc < 0,6$). Les CDS qui ne sont ni '*sure*' ni '*probable*' sont '*improbable*' (elles seront éliminées simplement sur des critères de longueur et/ou de Pc). Ces valeurs ont été choisies dans le cadre d'une annotation fine qui conviendrait à tout type de génome nouvellement séquencé (valeurs peu sévères). Dans le cadre d'une réannotation, on utilise plutôt les jeux de paramètres définis par la validation automatique et adapté à la richesse en G+C (valeurs plus sévères : *sure_L* = 60 pb, *sure_Pc*2 = 0,7, *prob_L*2 = 150 pb et *prob_Pc* = 0,4).
- *sure_ss_I* est égal à 10% si la CDS incluse est sur le même brin que la CDS la plus longue et *sure_os_I* est égal à 40% si elle est sur le brin complémentaire. Dans un cas d'inclusion entre deux CDS '*sure*' sur le même brin, la CDS incluse sera conservée si sa longueur représente au moins 10% de celle de la CDS la plus longue. Cette valeur a été déterminée d'après l'étude des décalages du cadre de lecture authentiques identifiés chez *B. subtilis*.
- Nous n'autorisons pas les chevauchements trop importants entre une CDS '*sure*' et une CDS '*probable*' en sens contraire (nous éliminons la CDS '*probable*' et bien entendu nous conservons la CDS '*sure*'). Nous avons choisi la limite supérieure de 10% pour *sure_prob_os_O* afin de prendre en compte la possibilité de problèmes persistants de position prématurée du codon d'initiation des CDS (les CDS alors trop longues peuvent générer des chevauchements artefactuels).
- *prob_glob_IO*, qui est le seuil de score maximum de recouvrement total d'une CDS '*probable*', est égal à 75%. Autrement dit, nous autorisons qu'une CDS '*probable*' ait un recouvrement total qui représente au maximum 75% de sa taille.

8.4 Optimisation automatique de la valeur des paramètres

Lors d'une collaboration avec G. Nuel du laboratoire Statistique des Génomes de l'université d'Evry, nous avons entrepris l'optimisation automatique de la valeur des paramètres *AMIGene* sur trois génomes pour lesquels nous possédons des annotations de référence (six expériences : valeurs sévères et peu sévères pour un génome G+C pauvre, moyen et riche).

La validation automatique des paramètres d'une méthode se déroule généralement en deux phases :

1. optimisation à partir des paramètres initiaux, en référence à deux[3] jeux de données
2. validation des paramètres optimaux en référence à un troisième jeu de données

Nous avons choisi la solution qui consiste à optimiser la valeur des paramètres d'*AMIGene* par rapport aux annotations de référence d'un génome et à valider ces valeurs par rapport aux annotations d'un autre génome dont le pourcentage en G+C est du même ordre de grandeur. Cette solution

[3]Les paramètres sont optimisés sur le premier jeu ; le second jeu permet de contrôler que les valeurs restent génériques et ne sont pas sur-ajustées sur le premier jeu (*over-fitting*).

évite de scinder aléatoirement les annotations de référence en plusieurs jeux (*e.g.* un jeu pour la phase d'optimisation et un jeu pour la phase de validation). De plus, elle permet de vérifier que les valeurs optimales ne sont pas génome spécifique mais sont applicables à n'importe quel génome dont le pourcentage en G+C est du même ordre de grandeur.

8.4.1 Caractéristiques des jeux de CDS de référence

Sachant que les caractéristiques des LS_CDS varient en fonction du pourcentage en G+C (nous avons déjà évoqué des différences au niveau des fréquences des codons d'initiation et de terminaison), nous avons décidé de valider les paramètres sur plusieurs génomes choisis en fonction de leur pourcentage en G+C et de la qualité de leurs annotations.

Nous avons donc optimisé les paramètres sur les trois génomes de référence (*B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv ; voir p. 31 et p. 212) puis nous avons validé les valeurs optimales sur trois autres génomes publiés plus récemment (*B. halodurans*, *E. coli* O157:H7 EDL933 et *M. tuberculosis* CDC1551) :

1. *B. subtilis* et *B. halodurans* [Takami & Horikoshi, 2000] sont deux espèces saprophytes du même genre (ordre des *Bacilles* et génome G+C pauvre). Le protéome de *B. halodurans* présente 75% de similitude avec celui de *B. subtilis*.
2. *E. coli* K-12 et *E. coli* O157:H7 EDL933 [Perna *et al.*, 2001] sont deux souches de la même espèce (*Entérobactéries* G+C moyen). *E. coli* O157:H7 EDL933 est une bactérie pathogène de l'homme. Les gènes du génome d'*E. coli* O157:H7 EDL933 ont été annotés à partir des résultats de prédiction de *GeneMark.hmm* [Lukashin & Borodovsky, 1998] et de recherche de similitude en utilisant Blast [Altschul *et al.*, 1990]. Les auteurs annoncent 75% de similitude nucléique entre les génomes d'*E. coli* O157:H7 EDL933 et d'*E. coli* K-12. Ils décrivent 177 îlots O (segments uniques à O157) de taille supérieure à 50 bp répartis sur 1,34 Mb et 234 îlots K (segments uniques à K12) répartis sur 0,53 Mb.
3. *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 [Fleischmann *et al.*, 2002] sont deux souches d'une même espèce (*Actinomycètes* G+C riches). Le génome de *M. tuberculosis* CDC1551 a été séquencé et annoté par le TIGR selon la procédure habituelle (voir p. 152). Les auteurs discutent des régions uniques de *M. tuberculosis* H37Rv par rapport à *M. tuberculosis* CDC1551 mais à aucun moment ils n'évoquent la différence entre les nombres de gènes annotés : 4304 CDS pour le génome de *M. tuberculosis* CDC1551 (4403 kb), et 3995 CDS pour le génome de *M. tuberculosis* H37Rv (4411 kb). Pourtant, il existe plus de 99% d'identité au niveau nucléique entre les génomes de *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 [Delcher *et al.*, 1999b].

Les annotations de ces six génomes ont été analysées et stockées dans PkGDB (voir p. 175). Nous avons construit des matrices de transition en fonction des classes de gènes d'usage des codons synonymes : trois pour *B. subtilis*, trois pour *B. halodurans*, trois pour *E. coli* O157:H7 EDL933 et quatre pour *M. tuberculosis* H37Rv (voir p. 220). Pour *M. tuberculosis* CDC1551, étant donné

que les deux souches sont très proches, nous avons utilisé les quatres matrices de *M. tuberculosis* H37Rv. A partir des matrices de transition, nous avons calculé les probabilités moyennes de codage (Pc) des CDS des banques stockées dans PkGDB et conservé la meilleure pour chaque CDS, ainsi que le numéro de matrice correspondant (voir le programme *GBK_max_Pc* p. 275). Dans le cadre du choix des paramètres d'une méthode de prédiction de CDS, il est important de connaître à la fois les caractéristiques de l'ensemble de départ (jeu de CDS totales d'un chromosome) et celles de l'ensemble d'arrivée (jeu de référence de CDS annotées). Le jeu de CDS totales d'un chromosome est l'ensemble des « vraies » CDS et des « fausses » CDS (prédites par *prokov_orf*). Généralement, on considère que l'ensemble des « vraies » CDS correspond au jeu de CDS annotées dans les banques. Ces caractéristiques sont, par exemple, la moyenne et la déviation standard d'un certain nombre de mesures comme la longueur des CDS, la probabilité moyenne de codage (Tab. 8.1 A p. 253). Ces connaissances aident à choisir des valeurs initiales pour les paramètres d'*AMIGene*.

Les caractéristiques générales du jeu de CDS totales définies par *prokov_orf* (LS_L > 60 pb) et du jeu de CDS annotées dans les banques (L > 60 pb) sont données dans le tableau 8.1 A p. 253. Plus le génome est G+C riche, moins il y a de CDS totales (la densité génique diminue) et plus elles sont longues (la couverture du génome augmente). Ceci peut s'expliquer par une fréquence en codons de terminaison moins importante chez les génomes G+C riches. Moins de 7% des CDS du jeu total sont de « vraies » CDS (*% CDS_nb ratio* ; Tab. 8.1 A p. 253). Les CDS totales ont une longueur moyenne de 210 pb et une Pc moyenne de 0,09 alors que les CDS annotées ont une longueur moyenne de 980 pb et une Pc moyenne de 0,83. Les critères de longueur et de Pc sont donc pertinents pour tenter de distinguer une « vraie » CDS d'une « fausse » CDS.

Comme le montre le tableau Tab. 8.1 A p. 253, la couverture génomique des CDS annotées de *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 est anormalement élevée (> 96%) par rapport aux autres génomes (< 90%). Au moins deux hypothèses peuvent expliquer cette couverture génomique anormalement élevée pour les CDS annotées :

1. Les protéines de *B. subtilis* sont en moyenne plus courtes que celles de *M. tuberculosis* H37Rv ; la différence de couverture génomique observée correspond donc à une réalité biologique.
2. Les protéines de *B. subtilis* ont en moyenne la même taille que celles de *M. tuberculosis* H37Rv ; la différence observée correspond donc à un artefact d'annotation.

Sous cette seconde hypothèse, les LS_CDS doivent être davantage réajustées dans leur partie 5′ non-codante chez *M. tuberculosis* H37Rv que chez *B. subtilis* pour éviter ce biais d'annotation. Sous l'hypothèse que le génome de référence de *M. tuberculosis* H37Rv est mieux annoté que celui de *M. tuberculosis* CDC1551, il est possible que la longueur moyenne des CDS annotées de *M. tuberculosis* CDC1551 soit inférieure à celle de *M. tuberculosis* H37Rv, parce que plus de faux-positifs de petite taille ont été annotés et font chuter la moyenne (et non pas parce que le codon d'initiation des CDS a été davantage réajusté chez *M. tuberculosis* CDC1551). Cette remarque est aussi valable pour le couple *E. coli* K-12 et *E. coli* O157:H7 EDL933.

Nous concluons de l'observation des jeux de CDS totales que les valeurs optimales des paramètres d'*AMIGene* seront différentes en fonction du pourcentage en G+C du génome analysé.

A

	Optimisation BACSU	Optimisation ECOLI	Optimisation MYCTU	Validation BACHD	Validation ECO57	Validation MYCTC
Gram	+	-	+	+	-	+
genome_L (pb)	4214630	4639221	4411532	4202353	5528445	4403836
% GC	43,5	50,8	65,6	43,7	50,5	65,6
tot_CDS_nb	63219	68848	59882	63169	81840	59769
AS_L Mean (pb)	176,44 ± 286,99	192,2 ± 269,79	265,91 ± 335,19	169,16 ± 245,57	191,61 ± 278,56	266,23 ± 333,15
AS_Pc Mean	0,092 ± 0,223	0,096 ± 0,221	0,094 ± 0,223	0,085 ± 0,219	0,097 ± 0,219	0,094 ± 0,222
gene density	1500	1484	1357	1503	1480	1357
% genome coverage	264,7	285,2	360,9	254,3	283,6	361,7
bk_CDS_nb	4105	4219	3995	4055	5342	4304
L Mean (pb)	912,25 ± 796,80	970,02±626,30	1069,22±801,87	899,77 ± 571,29	923,93 ± 696,22	1007,21±791,18
Pc Mean	0,840 ± 0,128	0,855 ± 0,121	0,790 ± 0,159	0,838 ± 0,126	0,821 ± 0,152	0,744 ± 0,205
gene density	97,4	90,9	90,5	96,5	96,6	97,8
% genome coverage	88,9	88,2	96,8	86,8	89,3	98,4
% CDS_nb ratio	6,5	6,1	6,7	6,4	6,5	7,2

B

	compute_Pc					AMI_filter_CDS															
sure-L	60																				
	valeurs testées	limites [	]	pas	valeur initiale	valeurs testées	limites [	]	pas	valeur initiale BACSU R2		BACSU R10		ECOLI R2		ECOLI R10		MYCTU R2		MYCTU R10	
climb-P	20	0,9	1	0,005	0,95					0,965				0,965		0,965		0,965		0,94	
diff-Pc	25	0,01	0,25	0,01	0,09					0,07				0,22		0,07		0,16		0,05	
sure-Pc1	20	0,05	1	0,05	0,7					0,9				0,75		0,7		0,75		0,75	
prob-L1	35	105	210	3	150					123				195		195		150		147	
prob-Pc					0,65	100	0,01	1	0,01	0,4	0,4	0,4	0,3	0,4	0,45	0,4	0,25	0,4	0,2	0,4	0,05
sure-Pc2					0,35	100	0,01	1	0,01	0,7	0,7	0,7	0,6	0,7	0,65	0,7	0,55	0,7	0,65	0,7	0,55
prob-L2					150	80	60	300	3	150	150	150	120	150	162	150	141	150	192	150	162
sure-ss-I					0,05	101	0	1	0,01	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
sure-os-I					0,3	101	0	1	0,01	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3
sure-prob-os-O					0,1	101	0	1	0,01	0,5	0,2	0,5	0,15	0,5	0,2	0,5	0,15	0,5	0,2	0,5	0,15
prob-glob-IO					0,8	126	0	1,25	0,01	0,8	0,7	0,8	0,9	0,8	0,7	0,8	0,9	0,8	0,7	0,8	0,9

(1) Optimisation des paramètres de compute_Pc

(2) Optimisation des paramètres de AMI_filter_CDS

TAB. 8.1 – Caractéristiques de jeux de CDS

A) Ce tableau présente les caractéristiques générales de longueur (L) et de probabilité moyenne de codage (Pc) de jeux de CDS. La première partie du tableau donne des caractéristiques sur les espèces bactériennes et leur génome. La deuxième partie caractérise les jeux totaux de CDS, issus de *prokov_orf* (AS ; AS_L > 60 pb ; [AGT]TG). La troisième partie caractérise les jeux publics de CDS annotées (L > 60 pb). La densité génique est le nombre de CDS par unité de longueur (Nb_CDS pour 100 kb). Le pourcentage de couverture du génome est le rapport de longueurs de l'ensemble des CDS sur le génome $((CDS\_nb * L * 100/genome\_L)$. Enfin, le pourcentage du rapport des nombres de CDS totales sur banque se calcule par $(bk\_CDS\_nb * 100)/tot\_CDS\_nb$.

B) Valeurs initiales pour l'optimisation automatique des onze paramètres de la stratégie de prédiction de CDS, *AMIGene*. L'optimisation se déroule en deux phases : optimisation des quatre paramètres de l'heuristique de réajustement de la position du codon d'initiation des CDS (*compute_Pc*), puis optimisation des sept paramètres de l'heuristique de filtrage des CDS (*AMI_filter_CDS*). On a indiqué en jaune foncé et en rose foncé les valeur initiales qui correspondent aux valeurs originales ajustées empiriquement.

Des histogrammes peuvent aussi aider à définir des valeurs initiales de seuils de longueur et de Pc. En comparant la distribution des longueurs des CDS des figures 8.6 A et B p. 255, on observe que *M. tuberculosis* H37Rv possède une proportion plus importante de longues CDS artefactuelles que *B. subtilis*. Les seuils de Pc sont faciles à choisir chez *B. subtilis* (*e.g. prob_Pc* ≈ 0,35 et *sure_Pc* ≈ 0,7) relativement à *M. tuberculosis* H37Rv (Fig. 8.6 C et D p. 255). Chez *M. tuberculosis* H37Rv, le seuil *prob_Pc* sera plus faible et les valeurs négatives artefactuelles nous gênent pour le choix de *sure_Pc* (elles sont probablement dues à des différences dans le réajustement de la position du codon d'initiation entre les CDS annotées et les AS_CDS).

8.4.2 Evaluation d'un jeu de paramètre

Principe

Afin d'évaluer les taux de prédiction de CDS, les grandeurs suivantes sont définies :

- Ngene est le nombre de « vrais » gènes (d'après les annotations de référence).
- Npred est le nombre de CDS prédites par *AMIGene* (pour un jeu de paramètres donné et avec des matrices de transition spécifiques des classes de gènes définies en fonction de l'utilisation des codons synonymes, voir p. 220).
- Nok est le nombre de CDS prédites correctement par *AMIGene* (*i.e.* ayant le même brin et la même position du codon de terminaison que la CDS de référence).
- Nmiss (Ngenes - Nok) est le nombre de gènes de référence manqués (ce sont les faux-négatifs).
- Nover (Npred - Nok) est le nombre de CDS prédites en trop par *AMIGene* (ce sont les faux-positifs).

Nous définissons alors les trois taux de prédictions de CDS suivant :

- La probabilité de prédire correctement un « vrai » gène vaut $P(pred \mid gene) = Nok/(Nok + Nmiss)$; autrement dit, c'est la fraction de « vrais » gènes correctement prédits. Cette quantité porte le nom de *sensibilité* (Sn) dans la littérature. Le taux de faux-négatifs peut défini tel que $FN = 1 - Sn$.
- La probabilité qu'une CDS prédite soit réellement un gène vaut $P(gene \mid pred) = Nok/(Nok + Nover)$; autrement dit, c'est la fraction des CDS prédites qui sont réellement des gènes. Cette quantité porte le nom de *spécificité* (Sp, ou de *sélectivité*) dans la littérature. Le taux de faux-positifs peut défini tel que $FP = 1 - Sp$.
- La probabilité conjointe d'être un gène et d'être prédit vaut $P(pred, gene) = P(pred \mid gene) * P(gene) = P(gene \mid pred) * P(pred) = Nok/(Nok + Nover + Nmiss)$; autrement dit, c'est la fraction des CDS totales qui sont correctes. Cette quantité porte le nom de *précision* dans la littérature.

Nous avons recherché la valeur des paramètres qui minimise la fonction de risque :

$$R\alpha = (\frac{\alpha}{\alpha + 1}) * FN + (\frac{1}{\alpha + 1}) * FP,$$

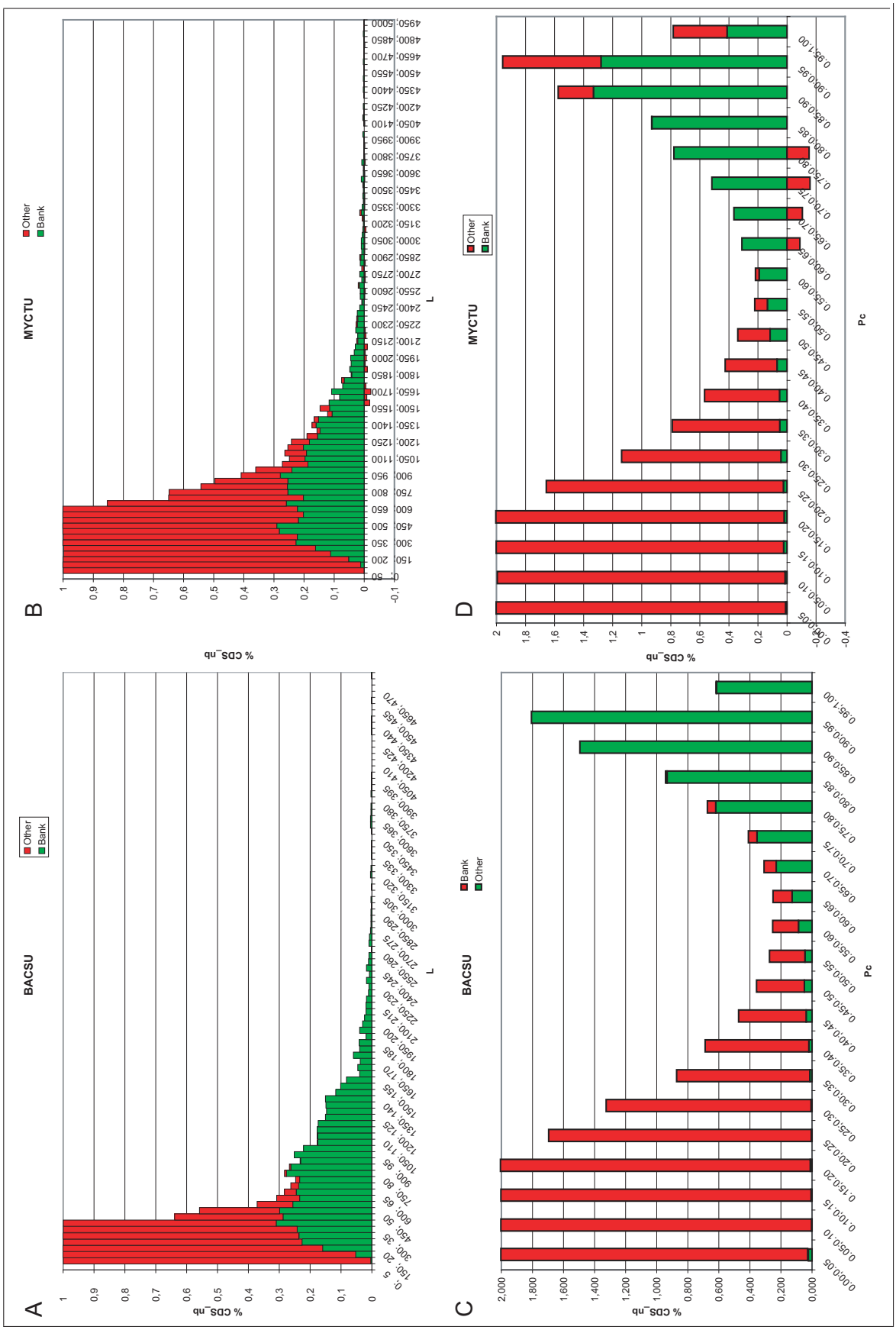


FIG. 8.6 – Histogrammes empilés des longueurs et des probabilités moyennes de codage des CDS
A) Distribution du pourcentage du nombre de CDS annotées empilé avec le pourcentage du nombre de CDS non annotées en fonction des classes de longueur chez *B. subtilis* (grande échelle).
B) Même qu'en A) chez *M. tuberculosis* H37Rv.
C) Distribution du pourcentage du nombre de CDS annotées empilé avec le pourcentage du nombre de CDS non annotées en fonction des classes de Pc chez *B. subtilis* (grande échelle).
D) Même qu'en C) chez *M. tuberculosis* H37Rv.

où FN est le taux de faux-négatifs, FP est le taux de faux-positifs et α est le facteur de pénalisation entre les prédictions de faux-négatifs et de faux-positifs ($\alpha/(\alpha+1)+1/(\alpha+1) = 1$ afin de normaliser le risque). La fonction de risque peut aussi s'écrire :

$$R\alpha = (\frac{\alpha}{\alpha+1}) * (1 - Sn) + (\frac{1}{\alpha+1}) * (1 - Sp),$$

où Sn et Sp sont respectivement la *sensibilité* et la *spécificité*. Nous avons décidé d'optimiser les paramètres dans deux conditions :

1. Une condition de paramétrage sévère (le nombre de faux-positifs est limité), pour laquelle il est deux fois plus pénalisant d'avoir un faux-négatif qu'un faux-positif ; cela correspond au poids $\alpha = 2$ de la fonction de risque $R2$ (*e.g.* jeu de paramètres *AMIGene* dans un projet de réannotation)
2. Une condition de paramétrage peu sévère (le nombre de faux-négatifs est limité), pour laquelle il est dix fois plus pénalisant d'avoir un faux-négatif qu'un faux-positif ; cela correspond à la minimisation de R10 (*e.g.* jeu de paramètres *AMIGene* dans un projet d'annotation).

Dans ces deux conditions, la sensibilité est plus pondérée que la spécificité.

L'étude des caractéristiques de jeux de CDS et l'optimisation empirique des paramètres nous ont permis de définir des valeurs initiales pour les jeux de paramètres à valider et nous n'avons donc pas eu besoin de tester plusieurs jeux de paramètres initiaux à tirer au hasard afin de diminuer la probabilité de tomber dans un minimum local.

Initialisation des paramètres de reconnaissance des CDS

Les conditions initiales que nous avons choisies pour l'optimisation automatique de la phase de reconnaissance des CDS correspondent au jeu de valeurs optimales expertes (valeurs originales des paramètres du réajustement du codon d'initiation, en jaune foncé ; Tab. 8.2 p. 257). Pour les trois génomes de référence, une diminution de la longueur moyenne des AS_CDS et une augmentation de la Pc moyenne est observée lorsque le *start* est réajusté automatiquement (valeurs initiales dans les conditions LS et AS ; Tab. 8.2 p. 257). Les écarts sont bien entendu plus marqués dans le cas du génome modèle G+C riche. De plus, pour un poids de 2, on s'aperçoit que ce jeu de valeurs initiales, testé sur trois génomes, donne le meilleur risque pour *B. subtilis*, avec une sensibilité de 98,44% et une spécificité de 92,01%. Pour un poids de 10, ce jeu de valeurs originales donne le meilleur risque pour *E. coli* K-12, avec une sensibilité de 98,63% et une spécificité de 91,41%.

Initialisation des paramètres de reconnaissance des CDS

Nous avons testé un certain nombre de conditions initiales (au moins deux pour chacun des trois génomes et pour chacune des deux conditions). Elles peuvent être séparées en deux catégories (Tab. 8.1 B p. 253) :

1. un jeu de valeurs optimales expertes (valeurs originales des paramètres de filtrage des CDS, en rose foncé),

	Climb-P	Diff-Pc	Sure-Pc	Prob-L	Sn	Sp	Pr	R	Mean total_CDS LS-L	AS-L	LS-Pc	AS-Pc
BACSU ori R2	0,950	0,09	0,7	150	98,44	92,01	90,69	3,703	177,90	176,89	0,09	0,091
BACSU R2	0,965	0,07	0,9	123	98,47	92,45	91,14	3,539		176,44		0,092
BACSU ori R10	0,950	0,09	0,7	150	98,44	92,01	90,69	2,144		176,89		0,091
BACSU R10	0,965	0,07	0,9	123	98,47	92,45	91,14	2,082		176,44		0,092
ECOLI ori R2	0,950	0,09	0,7	150	98,63	91,41	90,26	3,780	193,86	192,36	0,094	0,096
ECOLI R2	0,965	0,22	0,75	195	98,67	91,66	90,54	3,666		193,57		0,095
ECOLI ori R10	0,950	0,09	0,7	150	98,63	91,41	90,26	2,031		192,36		0,096
ECOLI R10	0,965	0,07	0,7	195	98,70	91,54	90,44	1,955		192,20		0,096
MYCTU ori R2	0,950	0,090	0,7	150	97,47	92,60	90,43	4,151	274,59	267,74	0,088	0,093
MYCTU R2	0,965	0,160	0,75	150	97,77	92,41	90,50	4,017		269,97		0,092
MYCTU ori	0,950	0,090	0,7	150	97,47	92,60	90,43	2,971		267,74		0,093
MYCTU R10	0,940	0,050	0,75	147	98,05	91,20	89,57	2,575		265,91		0,094

TAB. 8.2 – Synthèse de l'optimisation des paramètres de *compute_Pc*
Ce tableau synthétise les valeurs initiales et finales de l'optimisation des paramètres de l'étape de reconnaissance des CDS. Pour les étapes intermédiaires de l'optimisation, se reporter aux tableaux en annexe G.1 p. 416. Il présente aussi les caractéristiques générales de longueur (L) et de probabilité moyenne de codage (Pc) de jeux de CDS totales *Leftmost Start* (LS) ou *AMIGene Start* (AS) en fonction de différents jeux de valeurs de paramètres (originales, R2, R10).

2. des jeux de valeurs adaptées au poids et au génome (valeurs choisies par un expert, en rose clair)

Dans les exemples présentés dans le tableau 8.3 p. 263, nous observons que le jeu de valeurs adaptées au poids 2 et au génome de *B. subtilis* (en rose clair) donne le meilleur risque, avec une sensibilité de 98,08% et une spécificité de 92,07%. Pour un poids de 10, le jeu de valeurs optimales expertes (en rose foncé) donne le meilleur risque pour *E. coli* K-12, avec une sensibilité de 98,77% et une spécificité de 91,87%.

8.4.3 Optimisation automatique

Principe

Nous cherchons à optimiser la valeur des paramètres θ d'un algorithme. Autrement dit, nous cherchons le jeu de paramètres θ^* qui minimise la fonction de risque :

$$\theta^* = \arg \min_{\theta} R\alpha(\theta).$$

Dans notre cas, il existe quatre paramètres pour la phase de reconnaissance de CDS (*climb_P*, *diff_Pc*, *sure_Pc*1, *prob_L*1) et sept paramètres pour la phase de filtrage (*prob_Pc*, *sure_Pc*2, *prob_L*2, *sure_ss_I*, *sure_os_I*, *sure_prob_os_O*, *prob_glob_IO*), étant donné que nous ne cherchons pas à optimiser la longueur minimum d'une CDS '*sure*' ($sure\_L1 = sure\_L2 = 60$ pb). Nous sommes donc dans un espace réel à 11 dimensions : $\theta \in \mathbb{R}^{11}$. De plus, les propriétés mathématiques de la fonction sont inconnues. Si nous voulons explorer l'espace de toutes les valeurs possibles de θ, en testant 20 valeurs par paramètre et sachant qu'une évaluation prend cinq minutes, cela nous prendra alors $20^{11} * 5 \approx 10^{15}$ minutes, soit environ deux milliards d'années!! C'est pourquoi nous avons mis

en place une heuristique. L'optimisation automatique se déroule par minimisation successive dans chacune des onze dimensions. En pratique, nous minimisons la fonction de risque paramètre par paramètre, les autres paramètres étant fixés, en tournant au moins deux fois sur l'ensemble des paramètres (ce qui est généralement suffisant pour atteindre un équilibre). De cette manière, l'estimation du temps nécessaire à l'optimisation automatique n'est plus exponentielle en fonction du nombre de paramètres mais linéaire : $20 * 22 * 5 \approx 10^{15} \approx 2200$ minutes, soit environ 36 heures.

En pratique, nous allons analyser les résultats de deux expériences :

1. optimisation automatique des quatre paramètres de la phase de reconnaissance de CDS (les sept paramètres de la phase de filtrage étant fixés),
2. optimisation automatique des sept paramètres de la phase de filtrage (les quatre paramètres de la phase de reconnaissance étant fixés).

Expériences pour l'optimisation de la phase de reconnaissance des CDS

Lors de la validation automatique des paramètres de la stratégie *AMIGene*, nous étions surtout intéressés par la validation des paramètres de l'heuristique de filtrage (*AMI_filter_CDS* de la phase experte d'*AMIGene* Fig. 8.2 p. 242). En effet, la phase de reconnaissance ($\sim$ 5 min) est plus gourmande en temps de calcul que la phase de filtrage ($\sim$ 5 sec) ; nous aurions donc passé beaucoup de temps à valider les paramètres de l'heuristique de réajustement de la position du codon d'initiation des CDS, dont nous n'étions pas vraiment satisfaits (voir p. 268). Cependant, valider les paramètres d'*AMI_filter_CDS* sans avoir validé au préalable ceux de *compute_Pc* n'avait pas beaucoup de sens car la valeur de chacun des paramètres d'*AMI_filter_CDS* dépend de la qualité de la phase de reconnaissance. C'est pourquoi l'optimisation automatique des paramètres de la phase de reconnaissance a été moins fine que celle des paramètres de la phase de filtrage. Dans une première phase, nous fixons les paramètres d'*AMI_filter_CDS* et nous optimisons un à un, la valeur des quatres paramètres de *compute_Pc* ; dans une seconde phase, nous fixons les paramètres de *compute_Pc* et nous optimisons un à un, la valeur des sept paramètres d'*AMI_filter_CDS*.

Au cours de ce travail, nous avons cherché à optimiser les quatre paramètres de réajustement de la position du codon d'initiation des CDS de *compute_Pc* : *climb_P*, *diff_Pc*, *sure_Pc*1 et *prob_L*1 (Tab. 8.1 B p. 253). Nous ne pouvons tester qu'un nombre limité de valeurs pour chaque paramètre (environ 25) étant donné le temps de calcul de *compute_Pc*. Pour chaque génome (*B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv) et pour chaque condition (R2 et R10), nous avons itéré les minimisations successives dans chacune des quatre directions, jusqu'à ce que la variation du risque entre deux itérations successives soit inférieure à certain seuil (*e.g.* $(R_i - R_{i+1}) * 100)/R_i \leq 0,1\%$). Une itération teste environ $25 * 4 = 100$ conditions. Un équilibre est généralement atteint en moins de deux itérations. Ainsi, nous avons lancé six expériences, chaque expérience testant environ 200 conditions. Quelque soit l'expérience, nous avons utilisé les mêmes valeurs initiales qui correspondent aux valeurs originales ajustées empiriquement et aussi les mêmes valeurs pour les paramètres d'*AMI_filter_CDS* (Tab. 8.1 B p. 253). De plus, à chaque minimisation de la fonction

de risque, la stratégie *AMIGene* est exécutée complètement afin de calculer les taux de prédiction de CDS précédemment définis (on ne peut se contenter d'exécuter *compute_Pc*). Une expérience dure en moyenne 13 heures (200 * 242 / 3600).

Résultats de l'optimisation de la phase de reconnaissance

L'analyse des résultats du tableau 8.2 p. 257 permet de confirmer la nécessité de cette validation tout en remarquant que les valeurs originales étaient du bon ordre de grandeur. De plus, on vérifie que selon la richesse en G+C du génome et la valeur du facteur de pénalisation, les valeurs optimales des paramètres de *compute_Pc* ne sont pas les mêmes.

Après optimisation automatique des paramètres de *compute_Pc* pour *B. subtilis* et quel que soit le poids, on observe une diminution de la longueur moyenne des AS_CDS et une augmentation de leur Pc moyenne par rapport aux conditions initiales originales (Tab. 8.2 p. 257). Si l'optimisation de la position du codon d'initiation des CDS de *B. subtilis*n'entraîne pas une augmentation significative de la proportion de faux-positifs, on peut alors comprendre pourquoi on trouve le même jeu de valeurs optimales des quatre paramètres pour les facteurs de pénalisation 2 et 10. Si l'on compare les valeurs optimales aux valeurs initiales originales, on s'aperçoit que les valeurs *climb_P* et *sure_Pc*1 sont plus sévères et que les valeurs *diff_Pc* et *prob_L*1 sont moins sévères. Une augmentation de *climb_P* peut signifier que : (i) la qualité des matrices de transition a été améliorée ou (ii) la portion de la séquence sur laquelle un nouveau codon d'initiation est recherché, est allongée. Une diminution de *diff_Pc* signifie que le codon d'initiation est réajusté plus fréquemment. Augmenter *sure_Pc*1 signifie que pour réajuster le codon d'initiation des petites CDS de longueur comprise entre 60 et 123 pb, il faut qu'elles aient une AS_Pc supérieure ou égale à 0,9. Diminuer *prob_L*1 en revanche signifie que si une CDS n'a pas une AS_Pc supérieure ou égale à 0,9, il suffira qu'elle ait une AS_L supérieure à 123 pb pour que la position du codon d'initiation soit réajustée (Fig. 8.3 A p. 244).

Pour *E. coli* K-12 et *M. tuberculosis* H37Rv, on observe que les résultats obtenus dans les conditions originales sont intermédiaires entre les conditions optimales R2 et R10 (Tab. 8.2 p. 257). Le génome d'*E. coli* K-12 possède une densité génique, une longueur moyenne de CDS totales et une couverture génomique intermédiaires entre celles de *B. subtilis* et *M. tuberculosis* H37Rv (8.1 A p. 253). On trouve des valeurs optimales plus sévères pour R2 que pour R10. Dans le cas où l'on ne pénalise que deux fois plus les faux-négatifs que les faux-positifs, on réajustera moins souvent le codon d'initiation afin d'éviter une augmentation significative de la proportion de faux-positifs (*diff_Pc*= 0,22). Quel que soit le poids, on observe chez *E. coli* K-12 un seuil de probabilité moyenne de codage pour être une CDS '*sure*' (*sure_Pc*1) moins important et un seuil de longueur pour être une CDS '*probable*' (*prob_L*1) plus important que chez *B. subtilis*.

Comme pour *E. coli* K-12, les valeurs optimales de *M. tuberculosis* H37Rv sont plus sévères pour R2 que pour R10 (il faut limiter la proportion de faux-positifs) ; mais quelque soit le poids, les valeurs optimales de *M. tuberculosis* H37Rv sont moins sévères que celle d'*E. coli* K-12. Ainsi, même s'il faut limiter la proportion de faux-positifs, les problèmes de réajustement du codon d'initiation sont

plus importants chez *M. tuberculosis* H37Rv que chez *E. coli* K-12 (on réajuste plus fréquemment la position du codon d'initiation chez *M. tuberculosis* H37Rv que chez *E. coli* K-12 ; Tab. 7.1 p. 213).

En conclusion, il semble que trois facteurs principaux influencent la valeur optimale des paramètres pour le réajustement de la position du codon d'initiation des CDS effectué automatiquement par *compute_Pc* :

1. la qualité des matrices de transition (*climb_P*),
2. le facteur de pénalisation (*diff_Pc*) et
3. la richesse en G+C du génome (*sure_Pc*1 et *prob_L*1).

Que le génome soit G+C riche ou G+C pauvre, la fréquence du réajustement du codon d'initiation est du même ordre de grandeur : pour les trois génomes modèles, la fréquence de réajustement du codon d'initiation est supérieure à 40% (Tab. 7.1 p. 213). Si le réajustement du codon d'initiation est différent entre les génomes G+C riches et les génomes G+C pauvres, c'est principalement parce que la longueur moyenne de la partie $5'$ non-codante de la LS_CDS est plus importante chez les génomes G+C riches (la couverture génomique est 264,7 chez *B. subtilis* contre 360,1% chez *M. tuberculosis* H37Rv dans le tableau 8.1 A p. 253).

Expériences pour l'optimisation de la phase de filtrage des CDS

Au cours de ce travail, nous avons cherché à optimiser les sept paramètres de filtrage des CDS de *compute_Pc* : *prob_Pc*, *sure_Pc*2, *prob_L*2, *sure_ss_I*, *sure_os_I*, *sure_prob_os_O* et *prob_glob_IO* (Tab. 8.1 B p. 253). Nous pouvons tester un nombre important de valeurs pour chaque paramètre (environ cent) étant donné le temps de calcul d'*AMI_filter_CDS*. Pour chaque génome (*B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv) et pour chaque facteur de pénalisation, nous avons optimisé deux jeux de valeurs initiales : (i) un jeu de valeurs originales adaptée à tous les génomes procaryotes (utilisé pour R2 et R10) et (ii) un jeu de valeurs choisies en fonction des caractéristiques du génome et du facteur de pénalisation (Tab. 8.1 p. 253). Ainsi, nous avons lancé douze expériences, en procédant par itération des minimisations successives dans chacune des sept directions. Un équilibre est généralement atteint en moins de trois itérations ($((R_i - R_{i+1}) * 100)/R_i \leq 0,1\%$) et une itération teste environ $100 * 7 = 700$ conditions. A chaque minimisation de la fonction de risque, seule la phase *AMI_filter_CDS* de la stratégie *AMIGene* est exécutée afin de calculer les taux de prédiction de CDS (2100 * 5 / 3600 secondes par expérience soit environ trois heures). Elle utilise la liste de CDS issue de la phase de reconnaissance exécutée avec les paramètres optimisés selon le génome et le facteur de pénalisation (Tab. 8.1 B p. 253).

Résultats de l'optimisation de la phase de filtrage

Les résultats de l'optimisation automatique des paramètres d'*AMI_filter_CDS* sont présentés dans le tableau 8.3 A p. 263. De même que pour la phase de reconnaissance, leur analyse permet de confirmer la nécessité de cette validation tout en révélant que les valeurs originales étaient du bon

ordre de grandeur. Par rapport aux résultats de l'*Article II* p. 274, nous avons recommencé l'optimisation automatique de *M. tuberculosis* H37Rv en utilisant cette fois quatre matrices de transition au lieu de trois (voir p. 197). La phase d'optimisation de la valeur des paramètres d'*AMI_filter_CDS* dans la condition R2 a conduit à une *sensibilité moyenne* de 97,5 et à une *spécificité moyenne* de 95,8. Pour R10, *AMI_filter_CDS* a une *sensibilité moyenne* de 98,7 et d'une *spécificité moyenne* de 90,3. La sensibilité est bien plus importante pour R10 que pour R2 et la spécificité est bien plus importante pour R2 que pour R10 (dans les deux conditions, la sensibilité est bien supérieure à la spécificité). On a donc bien une meilleure sensibilité dans la condition R10 et une meilleure spécificité dans la condition R2.

Pour le génome de *B. subtilis*, les valeurs obtenues pour R2 montrent que *sure_ss_I* n'a aucun effet sur la diminution du risque (toutes les valeurs testées n'ont aucun effet sur le risque ; Tab. 8.3 A p. 263). Cela signifie qu'il n'y a pas d'inclusions entre les CDS '*sure*' qui sont dans le même sens, qu'elles appartiennent au jeu de CDS totales ou au jeu de CDS de référence. En revanche, il y a des inclusions entre les CDS '*sure*' sens contraire dans le jeu de CDS totales qui n'ont pas été annotées dans les banques de référence ; elles doivent donc être éliminées (toutes les valeurs de *sure_os_I* comprises entre 0,2 et 1 donnent le même jeu de CDS prédites par *AMI_filter_CDS* ; Tab. 8.3 A p. 263). La longueur de la CDS incluse qui doit être éliminée fait donc au maximum 20% de la CDS « englobante ». La valeur de *sure_prob_os_O* est nulle, ce qui signifie que, pour ce type de génome (G+C pauvre) et pour ce type de condition (R2), on ne doit autoriser aucun chevauchement entre une CDS '*sure*' et une CDS '*probable*' lorsqu'elles sont en sens contraire. L'ensemble de ces résultats suggère que cette séquence chromosomique contient peu de problèmes de décalage du cadre de lecture et/ou de réajustement du codon d'initiation Enfin, pour R10, les paramètres *prob_Pc*, *sure_Pc*, *prob_L*, *sure_prob_os_O* et *prob_glob_IO* ont des valeurs moins sévères que pour R2 alors que *sure_ss_I* et *sure_os_I* ont les mêmes valeurs.

Les résultats obtenus pour *E. coli* K-12 sont assez voisins de ceux de *B. subtilis* à ceci près que la précision est meilleure pour *E. coli* K-12 (R2 : Pr=94,26 et R10 : Pr=91,37 ; la spécificité est meilleure au détriment de la sensibilité ; Tab. 8.3 A p. 263). La valeur optimale de *prob_Pc* est plus sévère chez *E. coli* K-12 que chez *B. subtilis*, ce qui peut suggérer que les matrices de transition utilisées pour *E. coli* K-12 (construites par *AMIMat* à partir des classes de gènes voir p. 197 et Tab. 8.3 A p. 263) sont de meilleure qualité que celles de *B. subtilis*. On remarque aussi que les valeurs optimales de *prob_L* et *prob_glob_IO* sont aussi plus sévères chez *E. coli* K-12 que chez *B. subtilis*. Deux interprétations sont possibles :

1. Comme les matrices de transition d'*E. coli* K-12 sont meilleures que celles de *B. subtilis* et que *prob_Pc* est plus élevée, il y a une proportion moins importante de CDS '*probable*' chez *E. coli* K-12 que chez *B. subtilis* ; on peut donc se permettre d'être plus sévère sur des seuils comme *prob_L* et *prob_glob_IO*.
2. Comme il y a une proportion plus importante de « fausses » CDS '*probable*' chez *E. coli* K-12 que chez *B. subtilis*, les paramètres comme *prob_L* et *prob_glob_IO* doivent être plus sévères.

Les résultats obtenus pour *M. tuberculosis* H37Rv se démarquent clairement de ceux de *B. subti-*

lis et d'*E. coli* K-12 : c'est pour ce génome que l'on obtient la précision la plus faible (R2 : Pr=93,06 et R10 : Pr=88,06 ; la sensibilité est aussi la plus faible ; Tab. 8.3 A p. 263). En effet, le biais de composition nucléotidique est plus marqué entre *M. tuberculosis* H37Rv et *B. subtilis* (et dans une moindre mesure entre *M. tuberculosis* H37Rv et *E. coli* K-12), qu'entre *B. subtilis* et *E. coli* K-12. Les paramètres de Pc *prob_Pc* et *sure_Pc* ont les valeurs les moins sévères, ce qui peut suggérer que les matrices de transition sont de moins bonne qualité. Les paramètres de recouvrement entre CDS '*sure*', *sure_ss_I* et *sure_os_I* sont les plus sévères car (i) il y a une proportion de CDS '*sure*' plus importante (*sure_Pc* étant moins sévère) ou (ii) il y a plus de cas de « fausse » CDS '*sure*' incluse. Le paramètre de recouvrement entre CDS '*probable*', *sure_prob_os_O* est le moins sévère car (i) il y a plus de cas de « vrai » chevauchement entre une CDS '*sure*' et une CDS '*probable*' qui sont en sens contraire ou (ii) il persiste plus de problèmes de position du codon d'initiation pas suffisamment réajusté. Enfin, *prob_L* a la valeur optimale la plus élevée car (i) il y a une proportion plus importante de longues CDS artefactuelles chez *M. tuberculosis* H37Rv et/ou (ii) les matrices de transition de moins bonne qualité génèrent plus de bruit au niveau du non-codant.

En conclusion, quel que soit le génome étudié, il semble que la valeur optimale de certains paramètres tels que *prob_Pc* ou *sure_Pc* dépendrait plus de la qualité de matrices de transition, tandis que la valeur d'autres paramètres tels que *prob_L* ou *sure_prob_os_O* dépendrait plus du contenu en G+C du génome. Lorsque l'on annote un génome supposé contenir de nombreux décalages du cadre de lecture, tels que *Y. pestis* CO92, *R. prowazekii* ou *Mycobacterium leprosis*, il est recommandé de choisir une valeur peu sévère pour *sure_ss_I* afin de mettre en évidence les cas de *frameshift* compensé (*e.g.* 0,05). En revanche, les inclusions en sens contraire ne sont pas, à ma connaissance[4], annotées ; on peut donc choisir une valeur très sévère pour *sure_os_I* (*e.g.* 0,5 voire 1).

8.5 Validation automatique des paramètres optimisés

Les valeurs déterminées au cours des phases d'optimisation des paramètres de *compute_Pc* et d'*AMI_filter_CDS* de la stratégie *AMIGene* pour les trois génomes modèles et pour les deux facteurs de pénalisation ont ensuite été validées sur trois génomes publiés plus récemment et ayant un contenu en G+C similaire aux génomes de référence (Tab. 8.1 A p. 253) : *B. halodurans*, *E. coli* O157:H7 EDL933 et *M. tuberculosis* CDC1551.

8.5.1 Expériences

Pour les trois génomes *B. halodurans*, *E. coli* O157:H7 EDL933 et *M. tuberculosis* CDC1551 et dans les deux conditions R2 et R10, nous avons exécuté le programme *AMIGene* avec des matrices de transition spécifiques des classes de gènes d'usage des codons synonymes et avec les 11 paramètres optimisés. Nous avons donc lancé six expériences : exécution du programme *AMIGene*, comparaison

[4] A l'exception d'une IS qui pourrait s'insérer en sens inverse sans fragmenter la CDS. Ainsi, dans ce cas particulier, la CDS incluse en sens inverse devrait être conservée.

	prob-Pc	sure-Pc	prob-L	sure-ss-I	sure-os-I	sure-prob-os-O	prob-glob-IO	Sn	Sp	Pr	R	Nb CDSa
BACSU R2 ori	0,4	0,7	150	0,1	0,3	0,2	0,7	98,08	93,76	92,07	3,363	4603
BACSU R2	0,41	0,68	165	0,05	0,3	0	0,65	97,76	95,23	93,20	3,084	4214
BACSU R10 ori	0,3	0,6	120	0,1	0,3	0,15	0,9	99,03	86,88	86,14	2,078	4679
BACSU R10	0,35	0,67	114	0,05	0,3	0,05	0,86	99,03	89,13	88,35	1,874	4561
ECOLI R2 ori	0,4	0,7	150	0,1	0,3	0,5	0,8	98,63	92,36	91,19	3,462	4505
ECOLI R2	0,47	0,7	189	0,03	0,3	0,16	0,6	97,77	96,33	94,26	2,708	4281
ECOLI R10 ori	0,4	0,7	150	0,1	0,3	0,5	0,8	98,77	91,87	90,82	1,860	4536
ECOLI R10	0,4	0,62	141	0,05	0,56	0	0,75	98,82	92,38	91,37	1,770	4513
MYCTU R2 ori	0,2	0,65	192	0,1	0,3	0,2	0,7	97,70	89,50	87,65	5,036	4361
MYCTU R2	0,36	0,66	225	0,14	0,31	0,16	0,8	96,92	95,89	93,06	3,423	4038
MYCTU R10 ori	0,05	0,55	162	0,1	0,3	0,15	0,9	98,72	73,51	72,82	3,569	5365
MYCTU R10	0,27	0,57	165	0,2	0,54	0,14	0,54	98,42	89,32	88,06	2,404	4402

	BACSU	ECOLI	MYCTU	BACHD	ECO57	MYCTC
Génome	4214	4639	4411	4202	5528	4403
Nb CDSb	4105	4219	3995	4055	5342	4304

	prob-Pc	sure-Pc	prob-L	sure-ss-I	sure-os-I	sure-prob-os-O	prob-glob-IO	Sn	Sp	Pr	R	Nb CDSa
BACHD R2	0,41	0,68	165	0,05	0,3	0	0,65	97,04	95,91	93,18	3,338	4103
ECO57 R2	0,47	0,7	189	0,03	0,3	0,16	0,6	95,15	97,34	92,74	4,119	5222
MYCTC R2	0,36	0,66	225	0,14	0,31	0,16	0,8	90,17	95,59	86,57	8,022	4060
Moyenne	0,413	0,680	193	0,073	0,303	0,107	0,683	94,12	96,28	90,83		
BACHD R1(	0,35	0,67	114	0,05	0,3	0,05	0,86	98,55	89,70	88,53	2,259	4455
ECO57 R10	0,4	0,62	141	0,05	0,56	0	0,75	96,80	93,15	90,37	3,532	5551
MYCTC R1(	0,27	0,57	165	0,2	0,54	0,14	0,54	92,82	90,26	84,37	7,412	4426
Moyenne	0,340	0,620	140	0,100	0,467	0,063	0,717	96,06	91,04	87,76		

TAB. 8.3 – Synthèse de l'optimisation des paramètres d'*AMI_filter_CDS* chez *B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv

A) Ce tableau synthétise les valeurs initiales et finales de l'optimisation des paramètres de l'étape de filtrage des CDS.

B) La taille des génomes est donnée en kb.

C) Les résultats de la validation automatique des paramètres d'*AMI_filter_CDS* chez *B. halodurans*, *E. coli* O157:H7 EDL933 et *M. tuberculosis* CDC1551. Abréviations : Sn, pourcentage de sensibilité d'*AMI_filter_CDS* ; Sp pourcentage de spécificité ; R, fonction de risque à minimiser pour trouver les valeurs optimales, CDSa pour les CDS d'*AMIGene* et CDSb pour les CDS de référence.

Pour les étapes intermédiaires de l'optimisation ainsi que pour des résultats supplémentaires, se reporter aux tableaux en annexe G.3 p. 416.

du jeu de CDS prédites par *AMIGene* dans les conditions optimales et du jeu de CDS annotées dans les banques, et enfin calcul des taux de prédiction de CDS.

8.5.2 Résultats

Les résultats de la validation automatique des paramètres de la stratégie de prédiction CDS *AMIGene* sont présentés dans le tableau 8.3 C p. 263. Par rapport aux résultats de l'*Article II* p. 274, nous avons recommencé la validation automatique de *M. tuberculosis* CDC1551 en utilisant les quatre matrices de transition de *M. tuberculosis* H37Rv (et non pas les trois matrices de *M. tuberculosis* CDC1551). La phase de validation de la valeur des paramètres d'*AMIGene* dans la condition R2 a conduit à une *sensibilité moyenne* de 94,12 et à une *spécificité moyenne* de 96,28. Pour R10, *AMIGene* a une *sensibilité moyenne* de 96,06 et d'une *spécificité moyenne* de 91,04. Entre la phase d'optimisation des paramètres d'*AMIGene* et la phase de validation, on perd en sensibilité et on gagne en spécificité. C'est ainsi que pour R2, on se retrouve avec une spécificité moyenne supérieure à la sensibilité moyenne. En regardant les résultats plus attentivement, on s'aperçoit que la chute de sensibilité entre *B. subtilis* et *B. halodurans*, et entre *E. coli* K-12 et *E. coli* O157:H7 EDL933, est inférieure à 3%, alors qu'entre *M. tuberculosis* CDC1551 et *M. tuberculosis* H37Rv, elle est supérieure à 5%.

On observe un phénomène à priori surprenant : on s'attend à ce que la validation donne de meilleurs résultats sur des couples de génomes proches par rapport à des couples de génomes éloignés. On s'attend donc à ce que les résultats de validation soient meilleurs pour *M. tuberculosis* CDC1551, puis pour *E. coli* O157:H7 EDL933 et enfin pour *B. halodurans*. Or nous observons le contraire.

En revanche, ces résultats deviennent cohérents si on considère que plus une catégorie de génomes est facile à annoter, plus les annotations sont homogènes et meilleurs sont les résultats de validation. Les génomes A+T riches étant plus faciles à annoter que les génomes G+C riches, on s'attend alors à ce que les résultats de validation soient meilleurs pour *B. halodurans*, que pour *E. coli* O157:H7 EDL933, et meilleurs pour cette dernière que pour *M. tuberculosis* CDC1551.

8.5.3 Interprétation des résultats obtenus chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551

Nous avons décidé de vérifier l'hypothèse suivante : bien que les génomes de *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 soient quasiment identiques, il existe une hétérogénéité d'annotation à l'origine de la baisse de la sensibilité de prédiction de CDS quand on applique *AMIGene* au génome de *M. tuberculosis* CDC1551 avec le jeu de paramètres optimaux de *M. tuberculosis* H37Rv. Nous avons représenté les distributions de longueur et de probabilité moyenne de codage (Pc) des jeux de CDS totales et annotées chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 (Fig. 8.7 p. 266). Effectivement, nous n'observons quasiment pas de différences entre les distributions de longueur et de Pc des jeux de CDS totales chez *M. tuberculosis* H37Rv et *M. tuberculosis*

CDC1551. En revanche, nous observons une proportion de CDS annotées de longueur comprise entre 100 et 250 pb plus importante chez *M. tuberculosis* CDC1551 que chez *M. tuberculosis* H37Rv. Nous observons une proportion de CDS annotées de Pc comprise entre 0 et 0,7 plus importante chez *M. tuberculosis* CDC1551 que chez *M. tuberculosis* H37Rv, tandis qu'entre 0,7 et 1, elle est moins importante. Afin de démontrer que les distributions de longueur et de Pc sont similaires pour les jeux de CDS totales, et différentes pour les jeux de CDS annotées de *M. tuberculosis* CDC1551 et *M. tuberculosis* H37Rv, nous avons utilisé les tests statistiques de la plate-forme de méthodes d'analyse statistique, Statistica (Tab. 5.2 B p. 172). Pour chaque test, nous voulons comparer deux échantillons indépendants (*e.g.* comparer les distributions des longueurs du jeu de CDS totales de *M. tuberculosis* H37Rv et du jeu de CDS totales de *M. tuberculosis* CDC1551). Si la différence entre deux distributions ne suit pas une loi normale ($N(\mu, \sigma)$), on ne peut utiliser le test paramétrique du t-test (le test de Student compare deux distributions en comparant leur moyenne, ce qui ne suffit pas dans notre cas). Nous avons donc utilisé des tests non paramétriques : le u-test de Mann et Whitney et le test Kolmogorov-Smirnov (avec nos données, le runs-test de Wald et Wolfowitz ne donnait pas de résultats fiables ; annexe H p. 421).

En résumé, nous avons comparé trois caractères (longueur, Pc et G+C3) pour trois jeux de CDS (totales, annotées dans les banques et prédites par *AMIGene*) chez trois couples (*B. subtilis*-*B. halodurans*, *E. coli* K-12-*E. coli* O157:H7 EDL933 et *M. tuberculosis* H37Rv-*M. tuberculosis* CDC1551) avec deux tests statistiques non paramétriques (u-test et ks-test).

Les résultats présentés dans la figure 8.7 p. 266, montrent que les jeux de CDS totales et prédites par *AMIGene* sont plus différents entre *B. subtilis*–*B. halodurans* et *E. coli* K-12–*E. coli* O157:H7 EDL933, qu'entre *M. tuberculosis* H37Rv–*M. tuberculosis* CDC1551. Or la sensibilité de prédiction de CDS de la phase de validation automatique des paramètres d'*AMIGene* est plus faible pour *M. tuberculosis* CDC1551 que pour *B. halodurans* et *E. coli* O157:H7 EDL933. En effet, pour le couple *M. tuberculosis* CDC1551-*M. tuberculosis* H37Rv, on observe une homogénéité des longueurs et des Pc entre les jeux de CDS totales et de CDS prédites, en contradiction avec l'hétérogénéité observée entre les jeux de CDS annotées (Fig. 8.7 p. 266). De plus, si l'on filtre les différents jeux de CDS selon un critère de longueur ou de Pc, on arrive plus ou moins à homogénéiser les annotations de *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 (*e.g.* $Pc \geq 0,5$ ou $L > 300$; 8.7 p. 266). Quel que soit le critère choisi, nous n'arrivons pas à homogénéiser les distributions de Pc des CDS annotées chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551. Ce résultat est cohérent avec l'allure des distributions de la figure 8.7 p. 266 : à partir du seuil de 300 pb, les distributions de longueurs des CDS chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 paraissent similaires, en revanche, les distributions de Pc semblent différentes sur tout l'intervalle de valeurs. Bien entendu, si l'on recommence la validation automatique de *M. tuberculosis* CDC1551 sur les jeux de CDS prédites et annotées, filtrées selon leur Pc, la sensibilité de prédiction d'*AMIGene* augmente (résultats non présentés).

En conclusion de cette validation automatique des paramètres d'*AMIGene*, nous retiendrons que :

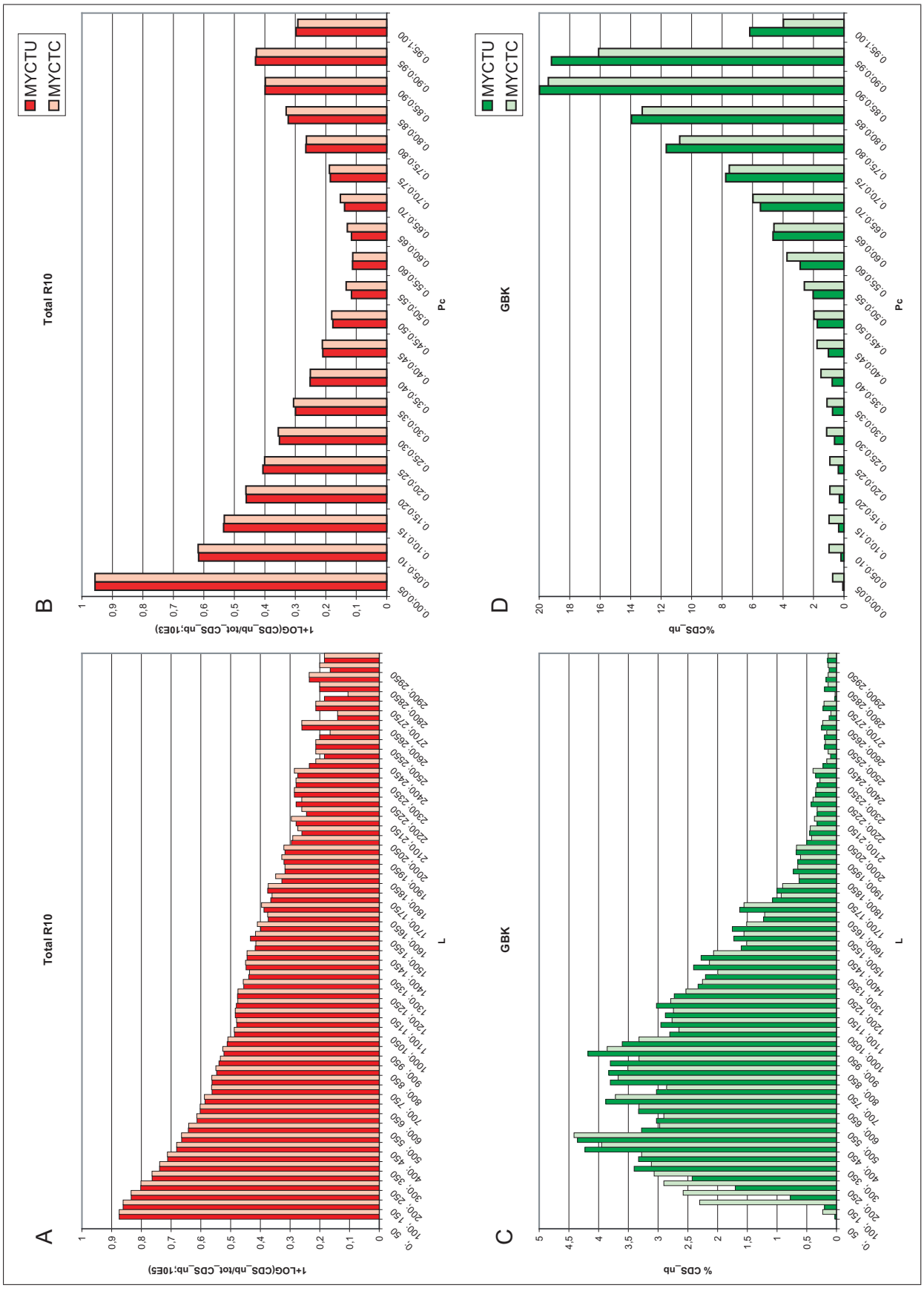


FIG. 8.7 – Hétérogénéité des annotations entre *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551

A) Distribution du LOG du rapport des nombres de CDS totales en fonction des classes de longueur des CDS chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551. L'intervalle de valeurs du pourcentage du nombre de CDS totales étant trop important (de 0,2 à 75%), nous avons dû adapter la mesure de manière à contraindre les valeurs dans un intervalle de 0 à 1.

B) Distribution du LOG du rapport des nombres de CDS totales en fonction des classes de Pc chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551.

C) Distribution du pourcentage du nombre de CDS annotées en fonction des classes de longueur chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551.

D) Distribution du pourcentage du nombre de CDS annotées en fonction des classes de Pc chez *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551.

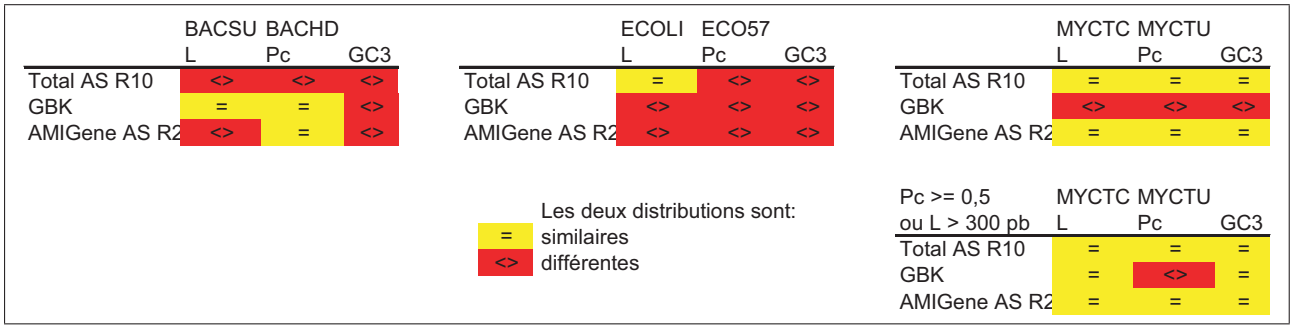


FIG. 8.8 – Comparaison des distributions de variables caractérisant les CDS entre des couples de génomes

Pour révéler la contradiction entre l'homogénéité des génomes *M. tuberculosis* H37Rv et *M. tuberculosis* CDC1551 et l'hétérogénéité de leurs annotations, nous avons comparé les distributions de longueur, de Pc et de G+C3 des jeux de CDS totales, annotées dans les banques (GBK) et prédites par *AMIGene* pour les trois couples de génomes *B. subtilis*-*B. halodurans*, *E. coli* K-12-*E. coli* O157:H7 EDL933 et *M. tuberculosis* H37Rv-*M. tuberculosis* CDC1551. Pour démontrer que deux distributions sont différentes, nous avons utilisé deux tests statistiques non paramétriques du logiciel Statistica : le u-test de Mann et Whitney et le test Kolmogorof-Smirnov (annexe H p. 421). Le test est significatif (*i.e.* les deux distributions sont différentes) lorsque la P-value est inférieure à 0,05. Il arrive que deux distributions soient similaires au niveau de leur valeur centrale, *i.e.* de leur somme des rangs (le u-test est non significatif) mais différentes au niveau de leur allure générale (le ks-test est significatif), ce qui nous mène dans ce cas à conclure qu'elles sont différentes (*e.g.* G+C3 des CDS totales entre *B. subtilis* et *B. halodurans*). Il arrive plus rarement que le u-test soit significatif alors que le ks-test est non significatif, ce qui nous mène dans ce cas à conclure que les distrubutions sont différentes (*e.g.* L des CDS de *B. halodurans* entre le jeu prédit par *AMIGene* et le jeu annoté dans les banques ; résultat non présenté). Finalement, il suffit qu'un de ces deux tests calcule une P-value significative pour que l'on puisse affirmer que les distribution sont différentes.

- Bien que nous ayons pris soin de choisir des annotations de référence, nous ne garantissons pas qu'il n'existe pas de biais d'annotation entre ces génomes (*e.g.* différence de longueurs moyennes des CDS).
- Il est recommandé de choisir consciencieusement la valeur des paramètres surtout lorsqu'il s'agit d'un génome G+C riche, car cette catégorie de génome est difficile à annoter.
- Il semble que les valeurs optimales définies pour un génome de référence G+C moyen, par exemple, soient applicables à un autre génome G+C moyen, mais nous ne garantissons pas que les valeurs optimales ne sont pas sur-ajustées par rapport aux annotations de référence.
- En pratique, nous utilisons généralement les valeurs optimales correspondant à un facteur de pénalisation des faux-négatifs égal à 10.
- De nombreux autres facteurs peuvent influencer la valeur optimales des paramètres d'*AMIGene* comme la qualité de la séquence chromosomique, des matrices de transition, la proportion de répétitions (*e.g.* IS, CDS fantômes), de régions remaniées récemment, de pseudogènes, etc.

8.6 Serveur web *AMIGene*

Au cours du développement d'un site Web par D. Vallenet, permettant de donner un libre accès au programme *AMIGene*[5]. Les fonctionalités de l'application Web et la méthode *AMIGene* propre-

[5] `http://www.genoscope.cns.fr/agc/tools/amigene/`

ment dite sont précisément décrits dans des pages d'aide de ce site Web. Les définitions relatives à l'ensemble des paramètres de la méthode sont accessibles à tout moment au niveau des ' *?*'. L'utilisateur choisit les paramètres d'*AMIGene* et les matrices construites au préalable en utilisant la stratégie experte *AMIMat*. Si les matrices ne sont pas disponibles pour la séquence procaryotique analysée, il choisit de construire automatiquement une matrice de transition (*prokov_learn* utilise un jeu de CDS d'une certaine longueur définies par *prokov_orf*, et du non-codant *shuffle*). Dans son état actuel, le site intègre aussi la méthode de détection de décalage de cadre de lecture *ProFED* [Médigue *et al.*, 1999b]. Cette méthode est en effet utilisée de façon systématique dans notre processus d'annotation de génomes bactériens. Les figures 7.2 A, B, C et D p. 204 présentent l'interface cartographique du site Web permettant de naviguer le long des six phases de la séquence nucléique pour explorer les CDS (rectangles rouges) prédites par *AMIGene*. L'ensemble des CDS et l'ensemble des protéines prédites sont téléchargeables dans un fichier au format fasta que l'utilisateur peut alors utiliser pour d'autres analyses ultérieures (*Article II* p. 274).

Etant donné le succès que rencontre le site *AMIGene*, J. Le Saux, a repris le code du programme *ProFED* afin d'y apporter des enrichissements, en particulier la détection de décalages du cadre de lecture à partir des résultats du programme de recherche de similitude dans les banques de séquences protéiques, Blast2x. Une seconde application Web, appelée *Micheck*[6], permet non seulement de rechercher des décalages du cadre de lecture avec cette nouvelle version de *ProFED*, à partir d'annotations déjà existantes dans les banques, et aussi de comparer ce jeu d'annotation au jeu de CDS prédites par *AMIGene* (CDS communes, uniques aux annotations et aux prédictions ; voir p. 281).

8.7 Originalités et limites actuelles du programme *AMIGene*

Il est intéressant de comparer les résultats d'*AMIGene* à ceux d'autres programmes de prédiction de CDS (Tab. 8.4 p. 269). Le jeu de CDS de référence correspond au jeu T défini par T.S. Larsen et A. Krogh (2042 CDS sûres[7] d'*E. coli* K-12 dont on a éliminé la redondance [Larsen & Krogh, 2003]). *AMIGene* exécuté sur le chromosome d'*E. coli* K-12 avec trois matrices et le jeu de paramètres correspondant à la condition R10 (Tab. 8.3 p. 263) permet d'obtenir 2029 CDS sur 2042 CDS du jeu T[8]. Comme attendu, la force d'*AMIGene* réside dans sa grande sensibilité aux dépends de sa spécificité. Cet écart est évidemment moins marqué pour $\alpha = 2$ que pour $\alpha = 10$. De plus, on remarque que les faux-positifs prédits par *AMIGene* dans les séquences aléatoires ont souvent été reconnus par la matrice de gènes de classe III (A+T riches).

A l'origine, la stratégie *Amiga* était exécutée au sein de la plate-forme Imagene en utilisant le programme *GeneMark* avec une seule matrice par chromosome réannoté (téléchargée à partir

[6] `http://www.genoscope.cns.fr/agc/tools/micheck/`

[7] Selon les auteurs, une CDS sûre a une longueur supérieure à 120 pb, son produit possède une similitude significative dans les banques de séquences protéiques et la fonction de ce produit est connue.

[8] C'est à confirmer car le fichier envoyé par T. S. Larsen contient des résultats de similitude. Il est donc difficile à exploiter (j'aurai préféré les coordonnées des CDS sur le chromosome).

Data set	EasyGene	Glim	Orpheus	Gm24	GmS	Gmhmm	Frame	AMIGene
T-% found	98,1(98,0)	98,3/98,4	96,5/95,6	89,8	96,3	97,1	96,1	99/99,4
Genome	4145	6827/5756	9333/7543	3552	4064	4230	4064	4286/4553
zero order	7	169/211	6761/5430	6	153	1459	0	294/1029
first order	7	545/723	6836/4804	13	241	830	0	450/1459
third order	1	2423/2694	6582/4817	43	659	866	1	597/1807

TAB. 8.4 – Performances d'*AMIGene* en comparaison avec d'autres programmes de prédiction de gènes

La sensibilité et la spécificité d'*AMIGene* sont comparées à celles d'autres programmes de prédiction de CDS. La colonne *AMIGene* a été ajoutée à la Table 2 de la publication *EasyGene* [Larsen & Krogh, 2003]. La partie supérieure présente le pourcentage de gènes trouvés (extrémité 3′ exacte) pour le jeu de 2042 CDS de grande confiance chez *E. coli* K-12 (sensibilité). La partie inférieure du tableau présente le nombre de faux-positifs prédits dans des séquences aléatoires générées par des chaînes de Markov d'ordre 0, 1 et 3 (spécificité). La première série de nombres de la colonne *AMIGene* correspond aux paramètres $\alpha = 2$ d'*E. coli* K-12 tandis que la seconde correspond à $\alpha = 10$ (TAB. 8.3 p. 263 ; les trois matrices d'*E. coli* K-12 ont été utilisées).

du site de *GeneMark*) ; les CDS prédites étaient des objets du SGBD d'Imagene (mais pas de PkGDB). La stratégie *Amiga*, écrite en *talk* et gourmande en ressources, nécessite de fragmenter le chromosome (l'exécution dure plusieurs heures sur un génome de 4 Mb).
Aujourd'hui, *AMIGene* est accompagné de la stratégie *AMIMat* qui permet de construire plusieurs matrices de transition spécifiques des classes de gènes d'usage des codons synonymes du chromosome réannoté (voir p. 197). Ces programmes, écrits en bash et en C, utilisent des modules du programme libre *Prokov* (l'exécution d'*AMIGene* dure quelques minutes sur un génome de 4 Mb). De plus, la table [AMIGene] a été ajoutée au modèle de PkGDB pour accueillir les CDS prédites (voir p. 185). Dans un proche avenir, nous aimerions pouvoir distribuer les sources d'*AMIGene* et intégrer la stratégie *AMIGene* dans le module *GenoAnnot* de la plate-forme *Genostar*.

Dans le cadre de la réannotation des 26 génomes procaryotes (*Article III* p. 289), nous avions choisi la valeur des paramètres d'*Amiga* empiriquement. Aujourd'hui, nous avons effectué une validation automatique de la valeur des paramètres d'*AMIGene*, qui permet de choisir entre six jeux de paramètres en fonction de la richesse en G+C du chromosome (pauvre, moyen ou riche), et du facteur de pénalisation des faux-négatifs (2 ou 10). De plus, les paramètres sont passés en argument du programme *AMIGene* en utilisant un fichier de configuration qui permet une gestion plus pratique de leur valeur (exécutions simultanées avec des jeux de paramètres différents et trace des différentes valeurs).

Initialement, la première étape de la phase de reconnaissance de CDS de la stratégie *Amiga* appelait une version modifiée du programme *SPOC* (*Simple Prokaryotic ORF and CDS*). Nous utilisions alors les trois codons d'initiation [ATG]TG. Ce programme a l'avantage, par rapport à *prokov_orf*, de pouvoir définir les CDS en combinant autant de signaux que l'on veut. En effet, *SPOC* utilise *SPat* pour la recherche de motifs, par exemple, de motifs de RBS. *SPOC* peut choisir le codon d'initiation en fonction de la présence d'un RBS. Nous avions modifié le programme *SPOC* pour qu'il créée des fragments de CDS aux extrémités des contigs dans les six phases de la

séquence[9].
Actuellement, nous utilisons *prokov_orf* qui définit les CDS uniquement sur la base de la présence des codons d'initiation et de terminaison (codon d'initiation le plus en 5', *Lefmost Start* (LS)). Nous utilisons les quatre codons d'initiation [ACGT]TG. Nous pourrions imaginer une nouvelle version de *SPOC* qui permettrait de définir les LS_CDS et proposerait une position alternative du codon d'initiation en combinant la présence simultanée de plusieurs signaux nucléiques de régulation de l'expression génique comme le terminateur, l'opérateur, le promoteur, le RBS. Au lieu d'utiliser uniquement *SPat*, *SPOC* appellerait aussi des programmes comme Petrin, RSAT, NNPP, RBSfinder (Tab. 3.1 p. 79) qui appartiennent à la nouvelle génération de programmes de découverte et de recherche de motifs. De plus, *SPOC* pourrait utiliser les fréquences d'apparition des codons d'initiation en fonction du pourcentage en G+C du génome étudié : par exemple, la fréquence d'ATG est égale à 0,78 chez les génomes G+C pauvres, 0,88 chez les génomes G+C moyens et 0,61 chez les génomes G+C riches (Tab. 7.1 p. 213). Néanmoins, il faut prendre garde aux biais d'annotation, par exemple, la fréquence d'apparition de 0,12 chez *B. subtilis* serait anormalement élevée, tandis que la valeur 0,02 chez *E. coli* K-12 serait plus proche de la réalité (K. Rudd communication personnelle).

Théoriquement, le calcul exact de la moyenne des probabilités de codage d'une CDS (Pc) nécessite que la probabilité de codage soit calculée à chaque position de la séquence en utilisant une fenêtre de taille $m+1$ (m ordre du modèle) et un pas de 1. En pratique, nous utilisons les valeurs par défaut de *GeneMark* (w=96 et s=12 pb). Le calcul de la variance des probabilités de codage d'une CDS est une mesure complémentaire au calcul de la Pc, qui permettrait de mieux caractériser la CDS. Une variance importante serait révélatrice d'un problème de démarrage trop en $5'$ ou d'un problème de profil atypique. Pour trancher entre ces deux problèmes, on peut alors étudier le nombre de fois où la courbe de probabilité franchit la limite de 0,5, le minimum et le maximum de la probabilité de codage le long de la CDS.

Le réajustement de la position du codon d'initiation des LS_CDS, tel qu'il est implémenté actuellement dans *AMIGene*, n'est pas suffisamment fiable, bien qu'il permette généralement de limiter les recouvrements entre CDS adjacentes. En particulier, certains codons d'initiation ne sont pas suffisamment réajustés tandis que d'autres le sont trop. Ceci est dû, respectivement, au problème des pics parasites (*i.e.* une montée artefactuelle de la courbe de codage avant le « vrai » codon d'initiation) et au problème du profil atypique (*i.e.* une courbe de probabilité de codage qui oscille sans jamais atteindre *climb_P*). Ainsi, pour la recherche de la position du « vrai » codon d'initiation, nous pouvons imaginer de combiner trois évidences :

1. la recherche par signal (terminateur, opérateur, promoteur, RBS)
2. la recherche par contenu (position du démarrage de la courbe de prédiction de codage selon un modèle statistique de séquences d'ADN par chaînes de Markov)

[9] Côté $5'$ du contig sur le brin direct : fragment sans start mais avec un stop. Sur le brin inverse : fragment avec un start mais sans stop. Côté $3'$ du contig sur le brin direct : fragment avec un start mais sans stop. Sur le brin inverse : fragment sans start mais avec un stop.

3. la recherche de similitude protéique en utilisant Blast2x

Les programmes *SHOW* et *EasyGene*, fondés sur les modèles de chaînes de Markov cachées, calculent aussi un score de codage et ajustent le codon d'initiation de chaque CDS (mêmes objectifs que la phase de reconnaissance de CDS d'*AMIGene* ; Fig. 8.2 p. 242). Il me semble que la probabilité moyenne de codage calculée par *AMIGene* est une mesure plus fine que leur score et qu'en revanche, *SHOW* et *EasyGene* réajustent mieux le codon d'initiation qu'*AMIGene*. Il est intéressant de réfléchir à la manière dont on pourrait associer ces programmes et/ou leurs résultats : soit en réconciliant les résultats après coup, soit en intégrant ces programmes au moment de la reconnaissance des CDS (avant les étapes de filtrage de CDS d'*AMIGene*).

La première étape de la phase de filtrage permet de constituer la liste des CDS '*sure*' et la liste des CDS '*probable*' selon des critères de longueur et de Pc (Fig. 8.3 C p. 244). Il serait intéressant d'affiner cette étape en explorant plus précisément la corrélation entre la longueur et la Pc de jeux de CDS de référence de génomes modèles, choisis de manière à couvrir l'intervalle du pourcentage en G+C. Par exemple, on peut essayer de définir l'équation de la courbe de tendance (*e.g.* régression linéaire) entre la longueur et la Pc sur un jeu de CDS dont la position du codon d'initiation a été vérifiée expérimentalement.

Enfin, trois points de l'heuristique de filtrage des CDS les plus probables sont ou vont être améliorés, afin d'être cohérent avec la recherche ultérieure de décalages du cadre de lecture (*ProFED*).

– Actuellement, dans *AMIGene* :

  1. Une CDS '*probable*' incluse dans une CDS '*sure*' est éliminée.
  2. Une CDS '*probable*' qui chevauche une CDS '*sure*' est éliminée si les deux CDS sont en sens inverse et que le pourcentage de chevauchement de la CDS '*probable*' est supérieur à *sure_prob_os_O*(*e.g.* 10%). *AMIGene* analyse les chevauchements '*probable*'-'*sure*' seulement s'ils concernent des CDS en sens contraire.
  3. Quand deux CDS '*probable*' (ou plus) se recouvrent avec un score total de recouvrement supérieur à *prob_glob_IO*, *AMIGene* conserve celle qui présente la plus forte Pc pour éviter de générer des trous d'annotation et élimine l'autre (Fig. 8.5 D p. 248).

– Dans une prochaine version d'*AMIGene*,

  1. Suivant le même principe que les chevauchements '*probable*'–'*sure*', on pourrait imaginer qu'une CDS '*probable*' qui est incluse dans une CDS '*sure*' dans le même sens soit conservée (*frameshift* compensé) et qu'en revanche, une CDS '*probable*' qui est incluse dans une CDS '*sure*' en sens contraire soit éliminée.
  2. Quand deux CDS '*probable*' (ou plus) se recouvrent avec un score total de recouvrement supérieur à *prob_glob_IO*, il semblerait judicieux, vu l'ambiguïté de la situation, de laisser le choix à l'annotateur. Face à un groupe de CDS '*probable*' qui doivent toutes être éliminées en raison de leurs recouvrements multiples, on regarde si elles sont plus nombreuses sur un brin ou sur l'autre : on élimine alors celles du brin où elles sont le moins

nombreuses et sinon on garde tout. Cette proposition ne gère pas le cas exceptionnel où une IS serait insérée en sens inverse dans une CDS alors fragmentée.

Pour conclure ce chapitre, le processus automatique de prédiction d'objets génomiques « de demain » doit être complet, intégré et doit faciliter le travail aux annotateurs (sans pour autant les remplacer).

On peut le voir à la manière d'*AMIGene*, où l'on essaie d'automatiser les différentes étapes de l'annotation manuelle. Nous avons vu par exemple, dans le cas du problème de réajustement de la position du codon d'initiation, que nous pouvions combiner au moins trois critères (*i.e.* recherche par signal, par contenu et extrinsèque) ; ces mêmes critères servent aussi à la détection de décalage du cadre de lecture. Nous pouvons donc aller plus loin : développer un processus qui combine un maximum de critères pour définir des objets génomiques et reconnaître les problèmes d'annotation comme le réajustement du codon d'initiation, le décalage du cadre de lecture, le profil atypique, et qui propose finalement un chemin probable à travers ces objets.
D'abord, lors de la phase d'*annotation syntaxique*, un maximum d'analyses éprouvées (*e.g.* recherches de motifs et de structures, recherches par contenu, recherches extrinsèques) sont enchaînées afin de définir un maximum d'objets génomiques de différents types (*e.g. tRNA*, *rRNA*, *RBS*, *CDS*, *terminator*, *similarity*).
Ensuite vient la phase d'*annotation fonctionnelle* : ces résultats sont synthétisés pour caractériser les objets (*e.g.* Pc, réajustement du codon d'initiation, CDS de classe I, II ou III, décalage du cadre de lecture). On attribue alors un statut de confiance multicritère aux objets génomiques. Par exemple, en fonction de la longueur de la CDS, de la Pc, de la présence de signaux de régulation, des résultats de similitude, on attribuera le statut rejeté, inconnu, putative, possible ou probable (Fig. 8.9) p. 273).
Puis, dans une phase d'*annotation relationnelle*, le type des objets est affiné et des objets génomiques de type complexe sont reconstruits (*e.g. fCDS*, *cCDS*, *traduction unit*, *operon*, *island*). Par exemple, on peut transformer deux CDS en fCDS si elles sont impliquées dans le même décalage du cadre de lecture (deux fCDS similaires à deux fragments du même polypeptide), et étendre, si nécessaire, ce groupe à d'autres fragments similaires pour définir précisément les bornes de la cCDS. La cCDS peut elle-même être impliquée dans un îlot de gènes excentriques, si on découvre par exemple un groupe de gènes de classe III colocalisés.
Enfin, la phase finale de filtrage des objets génomiques permet d'éliminer les faux-positifs en connaissance de toute l'information disponible. Dans le cas où l'on n'est pas capable de trancher automatiquement avec confiance, on garde tout pour laisser le soin à l'annotateur de décider ultérieurement. Ce processus est semi-automatique : l'utilisateur doit choisir des paramètres, vérifier le bon déroulement des différentes étapes. Il combine des sources d'évidence multiples qui sont alors analysées dans le contexte des objets génomiques.

On peut aussi voir le processus d'annotation automatique « de demain » à la manière de *SHOW* où l'on peut définir autant de types d'objets génomiques que l'on veut dans un seul modèle HMM (*e.g. RBS*, *rRNA*, *CDS*). De plus, chaque type d'objet peut avoir un nombre restreint de types de

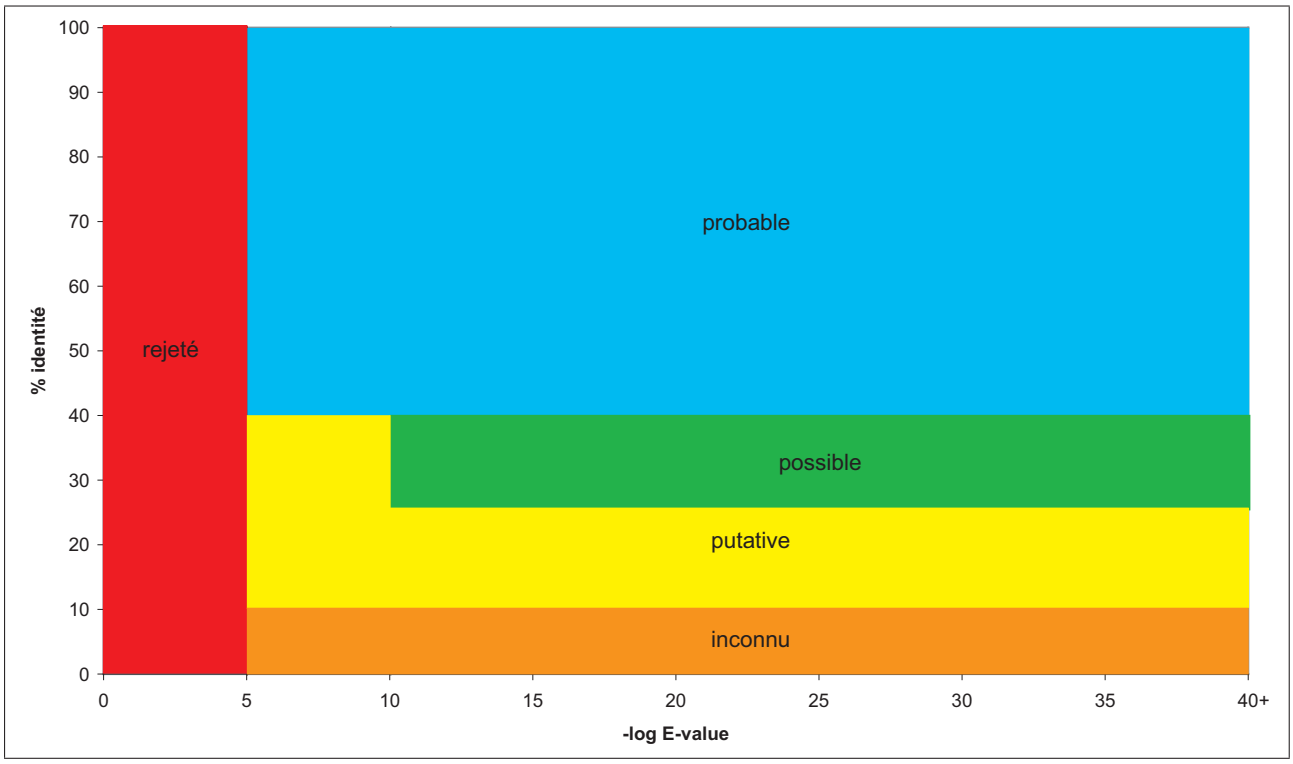


FIG. 8.9 – Exemple de schéma pour l'attribution d'un statut de similitude

Ce schéma permettant d'attribuer un statut de similitude aux CDS a été tiré des recommandations pour l'annotation de *Mycoplasma pulmonis* [Chambaud, 2000]. Les valeurs des cinq seuils (deux sur la E-value et trois seuils sur le pourcentage d'identité) ont été déterminées expérimentalement à partir des statistiques sur les résultats de Blastp. On peut imaginer le même genre de schéma basé sur le rapport des longueurs de l'alignement sur la plus petite des deux séquences et sur le pourcentage d'identité. On peut aussi imaginer le même genre de schéma à d'avantage de dimensions permettant d'attribuer un statut de confiance multicritère (*e.g.* longueur, Pc, similitude).

composition oligonucléotidique (*e.g.* pour le type *CDS* : *typical*, *highly expressed*, *A3+T3 rich*). La première partie du processus est complètement automatique, les paramètres sont transparents pour l'utilisateur et elle est fondée uniquement sur les caractéristiques intrinsèques de la séquence. Dans une second partie, on s'intéresse alors aux caractéristiques extrinsèques des CDS prédites et de leur produits.

Enfin, on peut le voir comme une réconciliation de plusieurs jeux de prédictions, chacun ayant été établi à partir d'une méthode de prédiction différente (*e.g. AMIGene*, *SHOW*, *EasyGene*
[Allen *et al.*, 2004]).

8.8 *Article II* : « AMIGene : Annotation of Microbial Genes »

S. Bocs, S. Cruveiller, D. Vallenet, G. Nuel, C. Médigue (2003)
Nucleic Acids Research, 31, 13.

Chapitre 9

Diagnostique sur de possibles erreurs de prédiction de CDS dans les banques INSD

L'Atelier de Génomique Comparative a pour objectif principal la comparaison des génomes et protéomes procaryotes. La génomique comparative repose sur la comparaison d'un jeu d'annotation d'objets génomiques (*e.g.* CDS) d'un génome G_1 et d'un jeu d'annotation d'objets génomiques d'un génome G_2. Logiquement, il suffit de récupérer les données disponibles dans les banques de séquences nucléotidiques. Cependant, tandis que le nombre de génomes complètement séquencés augmente continuellement, l'annotation de ces séquences déroge à la règle du « vite fait bien fait » [Devos & Valencia, 2001, Skovgaard *et al.*, 2001, Cruveiller *et al.*, 2003b, Iliopoulos *et al.*, 2003]. Ainsi, l'hétérogénéité des annotations rend difficile l'interprétation des résultats de comparaison (voir p. 65).

Différentes stratégies ont vu le jour pour fournir des jeux d'annotations de référence propres et « curés ».

1. Certains groupes d'annotation tiennent à jour leurs annotations. Par exemple, la mise à jour des annotations du génome de *M. pneumoniae* a été répercutée dans GenBank [Dandekar *et al.*, 2000] et celle des annotations des génomes de *B. subtilis* et *M. tuberculosis* H37Rv est en cours de transfert dans GenBank (à partir respectivement de la base SubtiList [Moszer *et al.*, 2002] et de la base TubercuList [Camus *et al.*, 2002]). Les annotations d'autres génomes sont uniquement mises à jour dans des bases spécialisées : par exemple, les annotations d'*E. coli* K-12 sont mises à jour dans la banque EcoGene [Rudd, 2000] et celles d'*H. pylori* 26695 et d'*H. pylori* J99 sont mises à jour dans la base PyloriGene [Boneca *et al.*, 2003].
2. Certains groupes de gestion de séquences biologiques mettent en place des projets multigénomes de *curation*, soit au niveau nucléique (*e.g. RefSeq* [Pruitt *et al.*, 2003]) soit au niveau protéique (*e.g.* HAMAP [Gattiker *et al.*, 2003]).

3. Certains groupes développent des méthodes et des plates-formes d'annotation, et réannotent les protéomes complets (*e.g.* GeneQuiz [Iliopoulos *et al.*, 2001], Pedant [Frishman *et al.*, 2003], COG [Natale *et al.*, 2000, Tatusov *et al.*, 2001]).

Les activités de l'Atelier de Génomique Comparative concernent ces trois aspects : premièrement, nous collaborons à des projets de réannotation de génomes procaryotes comme celui de *M. tuberculosis* H37Rv [Camus *et al.*, 2002] ; deuxièmement, nous collaborons à des projets de réannotation multigénomes comme HAMAP [Gattiker *et al.*, 2003] ; et troisièmement, nous avons mis en place un processus de réannotation de régions codantes à partir des données des banques, des résultats des méthodes *AMIMat*, *AMIGene* et *SWAN* ('*suspiciousBank*'–'*wrongBank*'–'*ambiguousAGC*'–'*newAGC*' ; [Bocs *et al.*, 2002, Bocs *et al.*, 2003]) et de la base PkGDB. Ainsi, le processus de réannotation est nécessaire à l'homogénéisation de la qualité et de la quantité des jeux d'annotations, pré-requis d'une génomique comparative pertinente (voir les résultats de réannotation dans le prochain chapitre p. 289).

9.1 Processus de réannotation

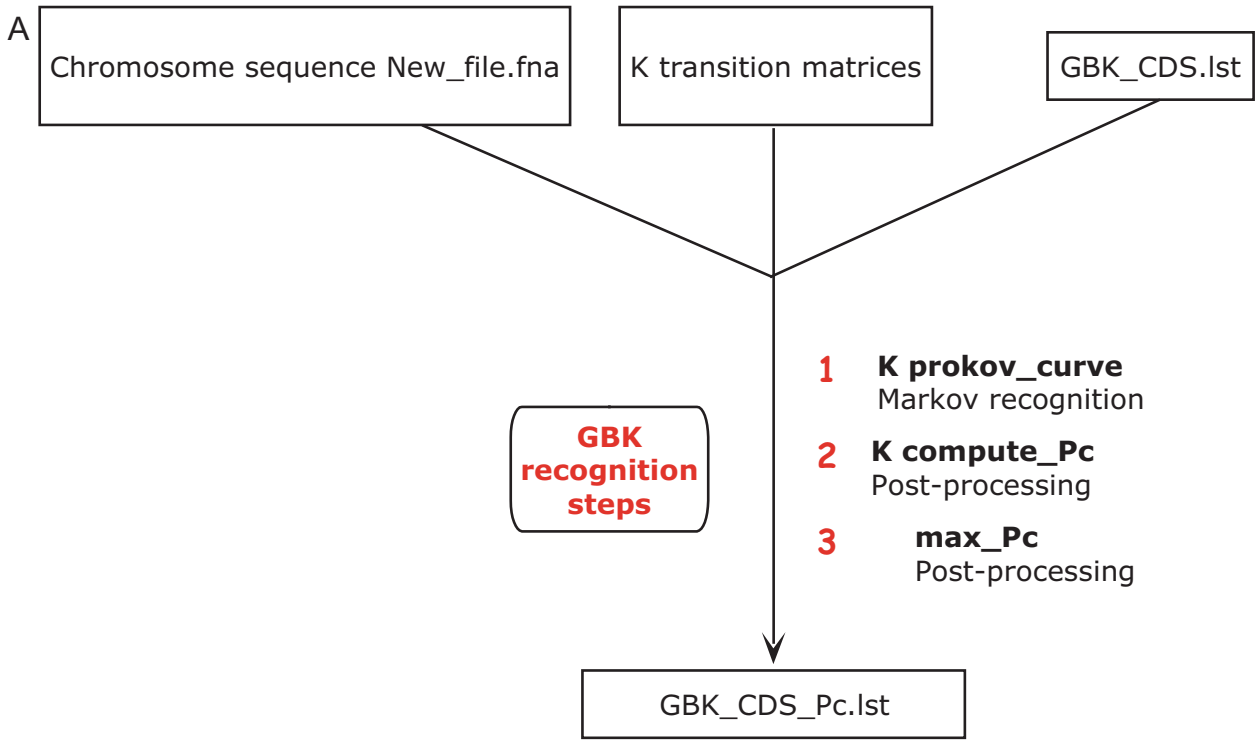
Le processus de réannotation de CDS des génomes procaryotes complets répertoriés dans les banques publiques, présenté dans cette section, est une version améliorée de celle qui est décrite dans la Figure 1 de l'*Article III* p. 289. Dans les deux versions, le principe reste le même : il s'agit de comparer deux jeux de CDS d'un même génome, l'un provenant des annotations des banques et l'autre des prédictions d'*AMIGene*.

9.1.1 Comparaison des prédictions d'*AMIGene* et des banques nucléiques

Ce processus de réannotation n'est actuellement appliqué qu'aux chromosomes complètement séquencés. La *première phase* du processus de réannotation comporte plusieurs étapes :

- Les annotations du chromosome du fichier *RefSeq* [Pruitt *et al.*, 2003] sont analysées par des programmes qui permettent de les formatter et de les stocker, de la table [Organism] à la table [Compare_Annotation] du modèle de données PkGDB (SGBDR MySQL ; voir p. 175). Cette dernière table contient notamment le jeu de CDS et de fragments de CDS des banques (fCDS ; voir p. 179).
- Les bornes anormales des CDS et des fCDS des banques de [Compare_Annotation] sont corrigées manuellement. L'annotateur expert utilise l'interface CompAnnotViewer qui facilite ces corrections (voir p. 183).
- Les k matrices de transition spécifiques de l'usage des codons synonymes du chromosome étudié sont construites à partir des CDS et fCDS des banques de [Compare_Annotation] et des séquences non-codantes natives (stratégie *AMIMat* ; voir p. 197). Cette stratégie nécessite l'expertise d'un annotateur.
- La probabilité moyenne de codage (Pc) des CDS et fCDS des banques est calculée par le programme *GBK_max_Pc*. Ce programme permet de calculer la meilleure Pc de chaque CDS

A



B

Abbreviation	Definition	Value
Sure_Pc	**Sure** coding **P**robability above which a CDS is interpreted as a sure CDS by the step combining two gene-finding intrinsic methods.	0.4 (0.6)
Sure_L	**Min**imum **L**ength of a **sure** CDS selected by the simple method searching for maximal CDSs.	60 bp
Prob_Pc	Between the *Prob_Pc* and the *Sure_Pc* **c**oding **P**robability thresholds, a CDS is interpreted as a **Prob**able CDS.	0.2 (0.3)
Prob_L	**Min**imum **L**ength of a **Pro**bable CDS selected.	150 (120) bp
Sure_ss_I	Minimum **I**nclusion percentage between two **sure** CDSs transcribed on the **s**ame **s**trand (Fig. 8.1 step 6).	10 %
Sure_os_I	Minimum **I**nclusion percentage between two **sure** CDSs transcribed on the **o**pposite **s**trands (Fig. 8.1 step 6).	30 (40) %
Sure_prob_O	Maximum **O**verlapping percentage between a **sure** and a **Prob**able CDSs, whatever their transcription orientation (Fig. 8.1 step 7). Seeing that the prediction of the correct translation initiation start remains an open problem, the value of this threshold is, at present,very high.	50 (10) %
Prob_glob_IO	Maximum **glob**al score, including both **I**nclusion and **O**verlapping situations, between a **Prob**able CDS and all the others probable CDSs being partially or completely in overlap with this CDS (Fig. 8.1 step 8). When a genomic region is devoid of annotation (*i.e.*, when all the overlapping probable CDSs are eliminated) we keep the CDS having the highest coding probability value.	80 (75)

FIG. 9.1 – Programme *GBK_max_Pc* et paramètres d'*AMIGene* pour la réannotation

A) Le programme *GBK_max_Pc*, utilisé lors de la *première phase* du processus de réannotation d'un chromosome complet procaryote, permet, à partir du chromosome, de k matrices de transition et d'une liste de CDS, de calculer les Pc et de conserver la meilleure ainsi que le numéro de matrice correspondant. Pour cela, il enchaîne les trois modules : *prokov_curve*, *compute_Pc* (sans réajustement de la position du codon d'initiation) et *max_Pc* (voir p. 207).

B) Ce tableau présente les valeurs choisies empiriquement des paramètres d'*AMIGene* (voir p. 246) utilisées dans la *deuxième phase* du processus de réannotation (*Article III* p. 289). Les paramètres *sure_L*, *sure_Pc*, *prob_Pc*, *prob_L*, *sure_ss_I*, *sure_os_I*, *sure_prob_O*et *prob_glob_IO* sont utilisés respectivement dans les étapes 1, 5, 5, 5, 6, 6, 7 et 8 décrites dans la figure 8.2 p. 242. A l'époque, nous utilisions une seule matrice par génome (pas d'analyse de l'usage des codons synonymes). Les nombres entre parenthèses indiquent les valeurs que nous prendrions aujourd'hui après quelques améliorations comme la stratégie *AMIMat* ou comme l'analyse des chevauchements entre CDS '*sure*' et '*probable*' seulement si elles sont en sens contraire. Plus précisément, suite à la validation automatique des paramètres d'*AMIGene*, nous utilisons trois jeux de paramètres en fonction du pourcentage en G+C du génome (faible, moyen ou riche ; TAB. 8.3 A R10 p. 263).

de la liste donnée en entrées, en fonction des différentes matrices d'utilisation des codons synonymes du chromosome étudié. C'est en fait une variante de la phase de reconnaissance d'*AMIGene* (Fig. 9.1 A p. 277 et voir p. 242). Les champs CA_Pc et CA_CU_matrix de la table [Compare_Annotation] sont alors mis à jour (voir p. 193).

Les CDS des banques sont donc définies et caractérisées par une position de début, une position de fin, une longueur (L), un sens, une phase, une Pc, un numéro de matrice correspondant à la meilleure Pc en fonction des k matrices de transition, un produit, une note, etc.

La *deuxième phase* du processus de réannotation consiste (i) à générer un second jeu de CDS pour le chromosome étudié à l'aide du programme *AMIGene* (*Article II* p. 274) qui utilise les k matrices de transition précédemment construites par la stratégie *AMIMat*, et (ii) à charger ces prédictions dans la table [AMIgene] de PkGDB (voir p. 185). Le choix de la valeur des paramètres d'*AMIGene* nécessite aussi une certaine expertise (Fig. 9.1 B p. 277). Ainsi, les CDS d'*AMIGene* sont définies par un identifiant, une position de début (le codon d'initiation le plus en 5′, *Leftmost Start* (LS), si la CDS est en sens direct), une position de codon d'initiation alternatif proposé par *AMIGene* (*AMIGene Start* (AS)), une position de fin et un sens. De plus, elles sont caractérisées par la phase, la longueur par rapport au *Leftmost Start* (LS_L), la longueur par rapport au *AMIGene Start* (AS_L), la LS_Pc, la AS_Pc et le numéro de la matrice de transition (correspondant à la meilleure AS_Pc conservée).

La *troisième phase* permet de comparer la liste des CDS annotées dans les banques et stockées dans la table [Compare_Annotation] et celle des CDS prédites par *AMIGene* et stockées dans la table [AMIGene] (voir p. 193). Au cours de cette comparaison, deux CDS sont dites identiques si la position du codon de terminaison est identique (on note cependant le nombre de CDS ayant un codon d'initiation différent). Le programme met à jour le champ CA_status ('*common*' ou '*uniqBank*') des CDS des banques. Enfin, il charge les CDS uniques à *AMIGene* dans [Compare_Annotation] (CA_status = '$uniqAGC$'). Au terme de cette comparaison, nous constituons trois listes (Tab. 10.1 p. 290) correspondant aux :

1. CDS communes aux annotations des banques de séquences et aux prédictions *AMIGene* (*Common Genes*, lst_CG),
2. CDS uniques aux annotations des banques de séquences (*Genes Not Found*, lst_GNF) et
3. CDS uniques aux prédictions *AMIGene* (*potential New Genes*, lst_NG).

9.1.2 Attribution d'un statut de similitude à certaines CDS uniques

La *quatrième phase* de la réannotation consiste à attribuer un statut de similitude à certaines des CDS uniques de lst_GNF et de lst_NG. En effet, il existe deux types d'erreurs d'annotation : la sur-annotation, qui correspond à l'annotation d'une « fausse » CDS (faux-positifs des banques), et la sous-annotation, qui correspond à l'oubli d'annotation d'une « vraie » CDS (faux-négatifs des banques). Nous voulons repérer des erreurs d'annotation flagrantes, par exemple, des faux-positifs dans la liste de CDS uniques aux annotations des banques (lst_GNF) et des faux-négatifs dans

la liste de CDS uniques aux prédictions *AMIGene* (*lst_NG*). C'est pourquoi nous cherchons à attribuer un statut de réannotation (et donc de similitude) aux :

1. CDS uniques aux annotations des banques et de faible Pc, pour la recherche de faux-positifs ('*improbable*', $lst\_GNF < prob\_Pc$, *e.g.* 0,4) et
2. CDS uniques aux prédictions *AMIGene* et de forte Pc, pour la recherche de faux-négatifs ('*sure*', $lst\_NG \geq sure\_Pc$, *e.g.* 0,7).

Les CDS des deux listes $lst\_GNF < prob\_Pc$ et $lst\_NG \geq sure\_Pc$ sont d'abord sélectionnées dans [Compare_annotation]. Puis les séquences nucléiques correspondantes sont extraites du chromosome par le module *lst2fna* (voir p. 207). Ensuite, le fichier de CDS est traduit en un fichier de polypeptides au format fasta (module *pwg_translate* est issu de la plate-forme Imagene). Enfin, les régions de basse complexité et les régions de répétitions de courte périodicité des séquences protéiques sont masquées respectivement par les programmes Seg et Xnu ([Wootton & Federhen, 1993, Claverie & States, 1993] et voir p. 319).

Le programme Blast2p de la plate-forme Biofacet [Glemet & Codani, 1997] est alors utilisé sur les séquences protéiques issues des listes $lst\_GNF < prob\_Pc$ et $lst\_NG \geq sure\_Pc$ afin de rechercher des similitudes dans la banque de séquences protéiques complète et non redondante SWALL [Boeckmann *et al.*, 2003]. Un autre programme permet d'analyser les résultats de Biofacet et d'attribuer un statut de similitude aux CDS de $lst\_GNF < prob\_Pc$ et de $lst\_NG \geq sure\_Pc$. Pour chaque séquence protéique, nous conservons, au maximum, les 20 premiers *hits* dont la E-value est inférieure à 10^{-3}.
Evidemment, le premier *hit* (ordonné selon la E-value) d'un polypeptide de $lst\_GNF < prob\_Pc$ correspond à une identité avec lui-même (100% d'identité sur toute la longueur, les produits des CDS annotées dans les banques de séquences nucléiques sont présents dans la SWALL). Chaque séquence protéique requête qui possède au moins deux *hits* est considérée comme ayant une similitude dans les banques (statut de similitude '*SIMTO*'). Celles qui n'en possèdent qu'un sont considérées comme n'ayant pas de similitude dans les banques ('*NOSIM*').
Pour les CDS de $lst\_NG \geq sure\_Pc$, un seul *hit* significatif suffit pour attribuer le statut '*SIMTO*' (et si elle n'a pas de hit, elle sera '*NOSIM*'). Ce statut de similitude est une caractéristique supplémentaire de la CDS, utile lors de l'attribution du statut de réannotation. A ce stade de l'analyse, nous connaissons donc pour chaque CDS : son identifiant, ses positions de début et de fin, sa longueur, sa phase, sa Pc, son numéro de matrice de transition et son statut de similitude.

9.1.3 Attribution d'un statut de réannotation à certaines CDS uniques

Le but de la *cinquième phase* est d'attribuer automatiquement un statut de réannotation aux CDS des listes $lst\_GNF < prob\_Pc$ et $lst\_NG \geq sure\_Pc$ grâce à la méthode *SWAN* qui combine plusieurs critères comme la longueur, la Pc, les recouvrements entre CDS et les résultats de similitudes (Figure 2 de l'*Article III* p. 289). Il s'agit d'être prudent afin de ne pas introduire de nouvelles erreurs en cherchant à en supprimer. Dans une procédure automatique d'attribution de

statut, deux approches sont possibles : (i) attribuer un statut à toutes les CDS avec un score de confiance associé à ce statut ou (ii) n'attribuer un statut que lorsque l'on a confiance en ce statut. Nous avons opté pour cette seconde solution. Les CDS pour lesquelles on ne peut attribuer un statut de réannotation automatique avec confiance pourront être analysées manuellement par un expert (elles ont explicitement le statut 'noStatus'). La méthode *SWAN* est constituée de deux heuristiques implémentées dans les programmes *GBK_sw.pl* et *AMI_an.pl* dont les algorithmes sont décrits en annexe I p. 423.

Le programme *GBK_sw.pl* attribue automatiquement un statut de réannotation à certaines CDS *uniques aux annotations des banques* et qui ont une *faible probabilité moyenne de codage* (Pc), en comparant les listes $lst\_GNF < prob\_Pc$ et $lst\_NG \geq sure\_Pc$ sur la base de critères multiples. Le pourcentage de recouvrement (*Inclusion–Overlap* (IO)) entre une CDS unique aux banques et de faible Pc (GNF '*improbable*'), et une CDS unique à *AMIGene* et de forte Pc (NG '*sure*') est le rapport des longueurs du recouvrement sur le GNF '*improbable*' :

$$IO(NGsur/GNFimpro)\% = \frac{L_{IO}}{L_{GNFimpro}} * 100.$$

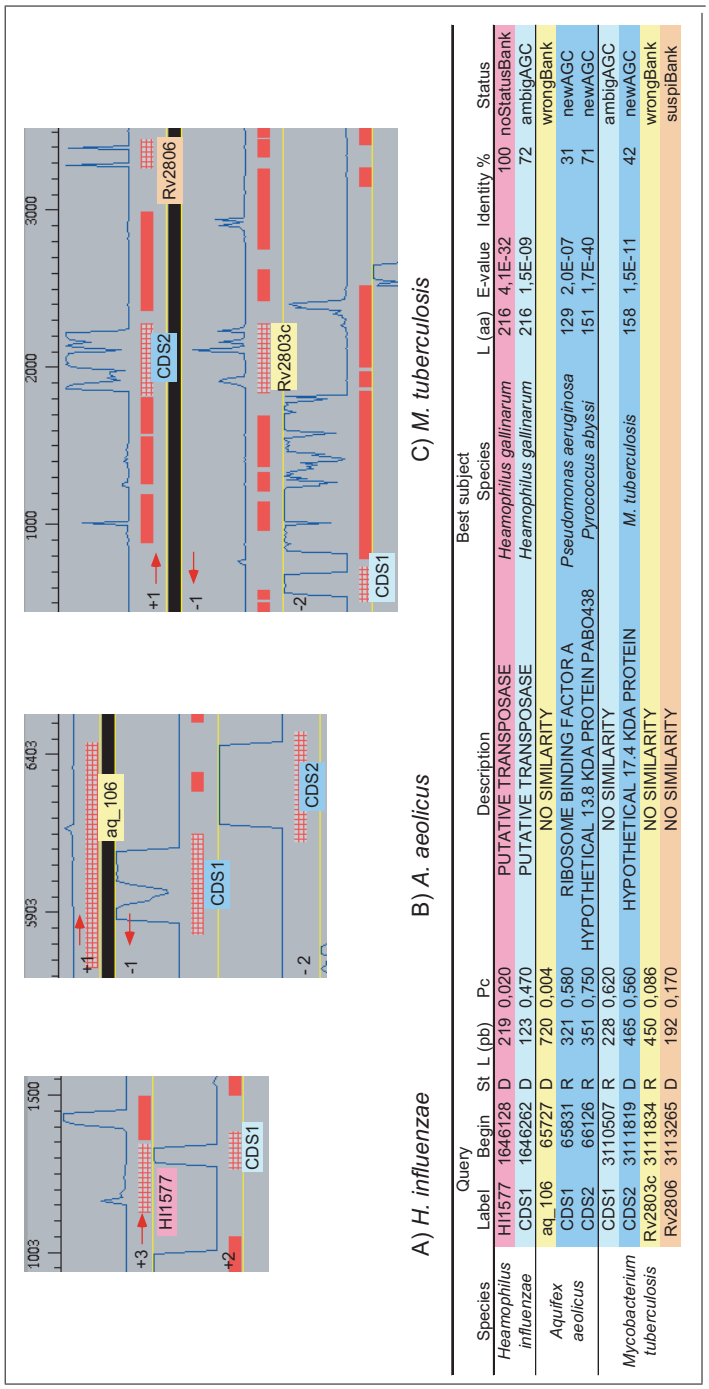
Le score de recouvrement total du GNF '*improbable*' est la somme de ses pourcentages de recouvrement avec les NG :

$$glob\_IO(GNFimpro) = \sum_i IO(NGsur_i/GNFimpro)\%.$$

Le statut de réannotation attribué aux CDS *uniques aux banques* et de *faible Pc* (*e.g.* $Pc < 0,4$) dépend des décisions suivantes :

1. 'noStatusBank' : La CDS (i) présente des similitudes dans les banques (*e.g.* $E-value \leq 10^{-3}$) *ou* (ii) est de grande taille (*e.g.* $L >$ 900 pb). Par exemple, la CDS HI1577 est qualifiée de '*noStatusBank*' dans la figure 9.2 A p. 281.
2. 'wrongBank' : La CDS de taille $\leq$ 900 pb ne présente pas de similitude (*e.g.* $E-value > 10^{-3}$) *et* (i) présente un recouvrement important (*e.g.* $glob\_IO(GNFimpro) \geq 50$) avec des CDS uniques à *AMIGene* et de forte Pc (*e.g.* $Pc \geq 0,7$) *ou* (ii) possède simplement une très faible Pc (*e.g.* $Pc < 0,2$) et une petite taille (*e.g.* $L <$ 300 pb). Par exemple, aq_106 a le statut '*wrongBank*' dans la figure 9.2 B p. 281.
3. 'suspiciousBank' : La CDS ne présente pas de similitude, est de taille $\leq$ 900 pb et n'a pas de recouvrement important avec les CDS uniques à *AMIGene* et de forte Pc (*e.g.* $glob\_IO(GNFimpro) < 50$) ; de plus, elle possède (i) une probabilité moyenne de codage comprise entre 0,2 et 0,4 *ou* (ii) une taille comprise entre 300 et 900 pb. Par exemple, Rv2806 a le statut '*suspiciousBank*' sur la figure 9.2 C p. 281.

Ainsi, le seuil du score global de recouvrement (*e.g.* $wrong\_glob\_IO = 50$) est décisif pour l'attribution du statut de réannotation '*wrongBank*' ou '*suspiciousBank*'. A l'issue du processus de réannotation, le statut des CDS uniques aux banques et de faible Pc est mis à jour au niveau du



Species	Query Label	Begin	St	L (pb)	Pc	Description	Best subject Species	L (aa)	E-value	Identity %	Status
Heamophilus influenzae	HI1577	1646128	D	219	0,020	PUTATIVE TRANSPOSASE	*Heamophilus gallinarum*	216	4,1E-32	100	noStatusBank
	CDS1	1646262	D	123	0,470	PUTATIVE TRANSPOSASE	*Heamophilus gallinarum*	216	1,5E-09	72	ambigAGC
Aquifex aeolicus	aq_106	65727	D	720	0,004	NO SIMILARITY					wrongBank
	CDS1	65831	R	321	0,580	RIBOSOME BINDING FACTOR A	*Pseudomonas aeruginosa*	129	2,0E-07	31	newAGC
	CDS2	66126	R	351	0,750	HYPOTHETICAL 13.8 KDA PROTEIN PABO438	*Pyrococcus abyssi*	151	1,7E-40	71	newAGC
Mycobacterium tuberculosis	CDS1	3110507	R	228	0,620	NO SIMILARITY					ambigAGC
	CDS2	3111819	D	465	0,560	HYPOTHETICAL 17.4 KDA PROTEIN	*M. tuberculosis*	158	1,5E-11	42	newAGC
	Rv2803c	3111834	R	450	0,086	NO SIMILARITY					wrongBank
	Rv2806	3113265	D	192	0,170	NO SIMILARITY					suspiBank

FIG. 9.2 – Exemples d'attribution automatique de statut de réannotation aux CDS uniques aux banques et à *AMIGene* : la méthode *SWAN*

Cette méthode est utilisée dans la *cinquième phase* du processus de réannotation d'un chromosome procaryote complet.

A) La CDS HI1577 est un GNF '*improbable*' qui a une similitude significative avec un fragment de CDS de transposase, elle n'a donc pas de statut ('*noStatusBank*'). Elle chevauche la CDS1 qui est un NG '*sure*' similaire à un autre fragment de la même transposase : le statut '*ambiguousAGC*' lui est donc attribué.

B) La GNF aq_106, '*improbable*' et '*NOSIM*' a un pourcentage de recouvrement de 44% inférieur à *wrong_glob_IO* avec la CDS1 '*sure*' et unique à *AMIGene* et un pourcentage de recouvrement de 44% avec la CDS2 '*sure*' et unique à *AMIGene*. Elle a un score de recouvrement total de 88% supérieur à *wrong_glob_IO*, qui lui vaudrait le statut '*wrongBank*' même si sa Pc avait été supérieure à *wrong_Pc*. Les CDS1 et CDS2 ont le statut '*newAGC*' (elles l'auraient même s'il n'y avait pas eu aq_106).

C) Les CDS1 et CDS2 sont inspectées pour l'attribution d'un statut car elles sont uniques aux prédictions d'*AMIGene* et ont une $Pc \geq 0,4$ (ancien *sure_Pc*). La CDS1 a le statut '*ambiguousAGC*' (et non pas '*newAGC*'), car elle ne présente pas de recouvrement avec un GNF mais elle est de petite taille ($L \leq new\_L$) et n'a pas de similitude significative avec les séquences protéiques de la SWALL. Nous avons attribué le statut '*wrongBank*' à Rv2803c parce qu'elle est '*improbable*' et '*NOSIM*' et qu'elle est en face de la CDS2 similaire à une CDS hypothétique des banques. La CDS2 est donc bien entendu '*newAGC*'. Rv2806 a le statut '*suspiciousBank*' (et non pas '*wrongBank*') car elle est '*improbable*' et '*NOSIM*' et ne présente pas de recouvrement avec un NG '*sure*' mais sa Pc est supérieure à *wrong_L*.

champs CA_status de la table [Compare_Annotation] de PkGDB. Finalement, le CA_status des CDS des banques peut prendre les valeurs :

- *'common'* pour les CDS communes aux annotations des banques et aux prédictions *AMIGene*,
- *'uniqBank'* pour les CDS uniques aux annotations des banques et de Pc moyenne ou forte (*e.g.* $Pc \geq 0,4$),
- *'noStatusBank'* pour CDS uniques aux banques, de faible Pc (GNF *'improbable'*) qui possèdent une similitude dans les banques (*'SIMTO'*) ou qui sont de taille > 900 pb.
- *'wrongBank' ou* 'suspiciousBank' pour les GNF *'improbable'*, *'NOSIM'* et de taille ≤ 900 pb.

Le programmme *AMI_an.pl* attribue automatiquement un statut de réannotation à certaines CDS *uniques aux prédictions AMIGene* et qui ont une *forte probabilité moyenne de codage* (Pc), en comparant les listes lst_GNF et $lst\_NG \geq sure\_Pc$, suivant le même principe que précédemment. Le pourcentage de recouvrement entre une CDS unique aux banques (GNF) et une CDS unique à *AMIGene* et de forte Pc (NG *'sure'*) est le rapport de la longueur du recouvrement sur la longueur du NG *'sure'* :

$$IO(GNF/NGsur)\% = \frac{L_{IO}}{L_{NGsur}} * 100.$$

Le statut de réannotation attribué aux CDS *uniques à AMIGene* de *forte Pc* (*e.g.* $Pc \geq 0,7$) dépend des décisions suivantes :

1. *'noStatusAGC'* : s'il existe un recouvrement significatif (*i.e.* $IO(GNF/NGsur)\% \geq 5\%$) (i) entre la CDS et une CDS unique aux banques et de Pc moyenne ou forte (*e.g.* $Pc \geq 0,4$) *ou* (ii) entre la CDS qui ne présente pas de similitude dans les banques (*e.g.* $E-value > 10^{-3}$) et une CDS unique aux banques, de faible Pc (*e.g.* $Pc < 0,4$) et qui présente une similitude (*e.g.* $E-value \leq 10^{-3}$).
2. *'ambiguousAGC'* : s'il existe (i) un recouvrement négligeable entre la CDS et une CDS unique aux banques et de Pc moyenne ou forte (*i.e.* $IO(GNFproOUsur/NGsur)\% < 5\%$) *ou* (ii) un recouvrement significatif entre la CDS qui présente une similitude, et une CDS unique aux banques, de faible Pc et ayant aussi une similitude.
3. *'newAGC'* : la CDS doit (i) n'avoir aucun recouvrement avec une CDS unique aux banques *ou* (ii) avoir un recouvrement avec une CDS unique aux banques de faible Pc et sans similitude *ou* (iii) avoir un recouvrement négligeable avec une CDS unique aux banques de faible Pc qui présente une similitude. Répondre à l'un de ces trois critères est une condition nécessaire mais non suffisante pour attribuer le statut *'newAGC'*. Il faut en plus que la CDS unique à *AMIGene* et de forte Pc (i) ait une taille significative (*e.g.* $L > 300$) *ou* (ii) présente une similitude significative (voir les CDS *'newAGC'* de la figure 9.2 B p. 281). Dans le cas contraire, elle aura le statut *'ambiguousAGC'* (voir la CDS *'ambiguousAGC'* de la figure 9.2 C p. 281).

Ainsi, le seuil du pourcentage de recouvrement (*e.g.* $new\_IO = 5\%$) est décisif pour l'attribution du statut *'noStatusAGC'*, *'ambiguousAGC'* ou *'newAGC'*. Pour le statut *'newAGC'*, aucun re-

couvrement n'est toléré entre la CDS unique à *AMIGene* et de forte Pc (NG '*sure*'), et une CDS unique aux banques et de Pc moyenne ou forte (GNF '*probable*' ou '*sure*') ; cela reste vrai quels que soient les autres recouvrements du NG '*sure*' et quel que soit le statut de similitude du GNF '*probable*' ou '*sure*'. Une différence importante entre les deux programmes est que, dans le cas de *GBK_sw.pl*, le score global de recouvrement est calculé pour le GNF '*improbable*', puis le statut de réannotation est attribué, alors qu'à chaque calcul de pourcentage de recouvrement entre un NG '*sure*' et un GNF, *AMI_an.pl* tente d'attribuer un statut. C'est pourquoi, dans ce second cas, une notion de priorité des statuts a été ajoutée : '*noStatusAGC*' est prioritaire sur '*ambiguousAGC*' qui est prioritaire sur '*newAGC*'. Finalement, le statut de réannotation des CDS *AMIGene* peut prendre les valeurs :

- *'*common*'*,
- *'*uniqAGC*'* pour les CDS uniques aux prédictions *AMIGene* et de Pc faible ou moyenne (*i.e.* $Pc < 0,7$),
- *'*noStatusAGC*'* s'il existe un recouvrement significatif entre la CDS unique à *AMIGene* et de forte Pc (NG '*sure*'), et un GNF qui est '*probable*' ou '*sure*', *ou* qui est '*improbable*' et '*SIMTO*' (et dans ce cas le NG doit en plus être '*NOSIM*'),
- *'*newAGC*'* *ou* *'*ambiguousAGC*'* pour les autres NG '*sure*'.

9.2 Intérêts et limites actuelles du processus de réannotation des génomes procaryotes

Ce travail de réannotation, dont les résultats sont présentés dans le chapitre suivant (voir p. 289), a donné lieu à plusieurs discussions au cours de diverses communications. Cela nous a permis de mieux mettre en lumière certaines de ses caractéristiques, ainsi que quelques lacunes. Nous nous proposons dans ce paragraphe de les présenter en introduisant les améliorations mises en œuvre ou à mettre en œuvre.

La réannotation des CDS des génomes procaryotes est une nécessité, comme le montre la quantité importante de projets de réannotation qui ont vu le jour ces dernières années et dont certains ont été présentés dans l'introduction de ce chapitre. L'originalité de notre processus de réannotation repose sur les points suivant :

- Il est semi-automatique (un expert doit vérifier le déroulement correct des cinq phases et prendre certaines décisions).
- Il est géré à l'aide d'un SGBDR.
- Il combine des critères intrinsèques et extrinsèques pour des prédictions syntaxiques et fonctionnelles.

Cette réannotation syntaxique de CDS permet de mettre en évidence des faux-négatifs (parfois confirmés par leurs recouvrements avec des faux-positifs), que les approches fondées uniquement sur la réannotation fonctionnelle des protéomes ne permettent de révéler.

Sachant qu'il n'existera pas de méthode parfaite capable de prédire toutes les « vraies » CDS (et

seulement elles) de tous les génomes procaryotes, le but de ce projet de réannotation n'est pas de critiquer les annotations originales en vantant les prédictions d'*AMIGene*, mais d'apporter de l'information biologique complémentaire et, si possible, pertinente. Ce travail permet d'homogénéiser les annotations, pré-requis essentiel de la génomique comparative (axe de recherche principal de notre groupe). Nous sommes en mesure de définir différents jeux de réannotation en fonction de l'objectif recherché :

- un jeu qui limiterait la proportion de faux-négatifs en regroupant par exemple les CDS ayant le statut '*common*', '*uniqBank*', '*noStatusBank*', '*suspiciousBank*', '*noStatusAGC*', '*ambiguousAGC*' ou '*newAGC*' (les CDS ayant le statut '*wrongBank*' ou '*uniqAGC*' sont donc rejetées),
- un jeu qui limiterait la proportion de faux-positifs en regroupant par exemple les CDS ayant le statut '*common*', '*uniqBank*', '*noStatusBank*' et '*newAGC*' (les CDS ayant le statut '*suspiciousBank*', '*wrongBank*', '*uniqAGC*', '*noStatusAGC*' ou '*ambiguousAGC*' sont donc rejetées).

Par ailleurs, ce travail intéresse la communauté scientifique, en particulier les groupes spécialistes d'un micro-organisme procaryote dans leur travail de mise à jour des annotations. Aussi, les cinq phases du processus de réannotation d'un chromosome procaryote ont-elles été améliorées.

En ce qui concerne la *première phase*, nous avons amélioré l'analyse, l'homogénéisation et le stockage des annotations des banques dans PkGDB (voir p. 179 [Vallenet, 2002, Jackson, 2002]). N'ayant pas prévu tous les cas possibles de format excentrique (voir p. 65), le « parseur » est perpétuellement en révision ce qui nécessite une certaine expertise. De plus, le format officiel[1] évolue aussi régulièrement (*e.g.* nouveaux attributs comme /locus_tag= ou /operon= pour les types d'objet *CDS* et *gene*, et vocabulaire contrôlé comme *synonym :* ou *synonyms :* dans l'attribut /note= du type *gene*). Par exemple, initialement, nous éliminions manuellement les CDS uniques aux banques et uniques à *AMIGene* détectées dans des régions de décalage du cadre de lecture annotées dans les banques, entre la *troisième phase* (comparaison des annotations) et *quatrième phase* (attribution d'un statut de similitude). Actuellement, nous avons mis en place une étape de correction manuelle des bornes des CDS anormales qui se déroule avant la comparaison des annotations (interface CompAnnotViewer adossée à PkGDB ; voir p. 183).
Nous avons aussi développé et validé la stratégie experte de construction de matrices de transition spécifiques de l'usage des codons synonymes des classes de gènes d'un chromosome pour la prédiction de phases codantes par chaînes de Markov (voir p. 234).

Pour la *deuxième phase*, qui consiste à prédire les phases codantes du chromosome en utilisant les différentes matrices construites lors de la première phase, nous avons développé, validé et amélioré le programme *AMIGene* (voir p. 268). Ainsi par exemple, l'utilisation de la matrice spécifique des gènes de classe III, permet de prédire les gènes de composition atypique (A+T riches ; Fig. 7.1 p. 198). Le choix de la valeur des paramètres nécessite aussi une certaine expertise (voir le choix empirique p. 246 et la validation automatique p. 250).

[1] `http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html`

La *troisième phase* de comparaison du jeu de CDS annotées dans les banques et du jeu de CDS prédite par *AMIGene* est fondée sur la comparaison de la position du dernier nucléotide du codon de terminaison d'une CDS des banques et d'une CDS *AMIGene* qui sont dans le même sens. Ce critère implique (i) que l'on compare des jeux d'annotation issus de la même version de la séquence chromosomique afin de ne pas décaler les positions du codon de terminaison des CDS et (ii) que les CDS se terminent toutes par un codon de terminaison. Ce dernier point entre normalement dans la définition de la CDS (séquence définie entre un codon d'initiation et le premier codon de terminaison rencontré en phase). Cependant, nous avons vu que les décalages du cadre de lecture entraînent des fragments de CDS qui respectent difficilement cette définition. En effet, le curateur a parfois du mal à trouver un codon de terminaison qui reflète les bornes des résultats de similitudes pour ces fragments de CDS. On pourrait donc imaginer de comparer les deux jeux d'annotation sur la base des résultats de similitudes (entre ces deux jeux), mais cette solution comporte aussi ses faiblesses. L'idéal serait peut-être, une fois de plus, de combiner les deux façons de faire.

Pour la *quatrième phase* d'attribution du statut de similitude '*SIMTO*' ou '*NOSIM*' aux CDS uniques aux annotations des banques et de faible Pc (GNF '*improbable*'), et aux CDS uniques aux prédictions *AMIGene* et de forte Pc (NG '*sure*'), le critère de décision fondé sur la E-value n'est certainement la meilleure solution. Pourrait être qualifié de similaire toute séquence protéique qui possède au moins un hit dont le pourcentage d'identité est supérieur à un certain seuil (*e.g.* 40%) et dont le rapport des longueurs de l'alignement sur la plus petite des deux séquences est supérieure à un autre seuil (*e.g.* 80%). C'est de cette manière que sont définis les couples d'orthologues de la table [GO_GO_CPD] de PkGDB (voir p. 190). Nous allons rechercher des similitudes de séquences contre la nouvelle base UniProt (successeur de la banque SWALL [Apweiler *et al.*, 2004]).

En ce qui concerne la *cinquième phase*, il est important de souligner que nous n'attribuons un statut de réannotation qu'à une petite proportion des CDS uniques aux annotations des banques (GNF) et des CDS uniques aux prédictions *AMIGene* (NG ; voir p. 295). Ainsi, nous sommes tout-à-fait conscients qu'une partie significative des GNF '*uniqBank*' ou '*noStatusBank*' sont des CDS avérées et qu'une partie significative de NG '*uniqAGC*' ou '*noStatusAGC*' sont des artefacts (faux-positifs). En revanche, il existe aussi des faux-positifs parmi les GNF '*uniqBank*' ou '*noStatusBank*' et des faux-négatifs parmi les NG '*uniqAGC*' ou '*noStatusAGC*' (Tab. 10.6 p. 320 et voir p. 302). L'ajustement des algorithmes *GBK_sw.pl* et *AMI_an.pl* et des paramètres s'est fait progressivement par essais–erreurs entre (i) des critiques soulevées par l'analyse des statuts de réannotation des CDS d'un génome et (ii) la modification en conséquence des programmes.
L'analyse des premiers résultats de la méthode *GBK_sw.pl* nous a mené à réajuster les critères d'attribution de chacun des statuts. Nous avions très peu de GNF '*wrongBank*' car les cas de recouvrement entre un GNF '*improbable*' et sans similitude, et un NG '*sure*' sont rares. Nous avons donc ajouté une possibilité : pour être '*wrongBank*', il suffit que le GNF sans similitude ait une très faible Pc et une petite taille. En revanche, nous avions de longues CDS qui avaient le statut '*suspiciousBank*'. Nous avons donc ajouté une contrainte : ne peuvent avoir le statut '*wrongBank*' ou '*suspiciousBank*' que les GNF '*improbable*' et sans similitude dont la longueur n'est pas trop

importante.
De même, lors de l'analyse des premiers résultats de la méthode *AMI_an.pl*, nous avions des petits NG '*sure*' et sans similitude qui avaient le statut '*newAGC*'. Nous avons donc introduit une condition supplémentaire : pour être '*newAGC*', il faut que le NG '*sure*' ait une longueur significative ou présente une similitude significative. Ainsi, certaines CDS qui devaient au départ avoir le statut '*newAGC*' vont devenir '*ambiguousAGC*' si elles sont de petite de taille et sans similitude.

Le paramétrage du processus de réannotation devrait être ajusté par une validation automatique. Nous devrions réannoter les chromosomes mais aussi les plasmides. Actuellement, lors de l'attribution d'un statut de réannotation à une CDS unique, seules les listes de CDS uniques sont comparées. Il semble plus judicieux de prendre en compte le contexte de la CDS unique dans sa globalité, en considérant par exemple les recouvrements et les résultats de similitudes des CDS adjacentes, qu'elles soient communes ou uniques. Cette modification devrait permettre d'améliorer la précision de la méthode *SWAN* et de confirmer ou de découvrir des décalages du cadre de lecture fragmentant les CDS. Comme toujours, l'analyse automatique ne peut se substituer entièrement à l'analyse manuelle par un expert : l'analyse automatique facilite simplement l'analyse manuelle ultérieure. L'analyse manuelle des statuts de réannotation attribués automatiquement aux CDS d'un génome est facilitée par l'utilisation d'une interface cartographique d'annotation telle que *MaGe* (Tab. 5.2 B p. 172). Dans le cas d'un génome pour lequel on voudrait faire une réannotation la plus exacte possible, il est alors conseillé d'analyser manuellement une à une toutes les CDS uniques aux annotations (GNF) et aux prédictions (NG).

Troisième partie

Prédictions biologiques

Chapitre 10

Réannotation de génomes bactériens complets

Ce chapitre résume et complète les travaux de réannotation *in silico* des génomes procaryotes complets présentés lors de plusieurs communications : poster [Bocs *et al.*, 2000], communications orales [Bocs, 2000, Médigue, 2000b, Médigue, 2001] et rapports [Médigue, 2000a, Vincent, 2001, Jackson, 2002]. Des exemples de réannotation de gènes procaryotes sont présentés dans l'*Article III* p. 289). L'*Article IV* p. 316 donne un exemple de validation expérimentale d'un nouveau gène prédit.

10.1 *Article III* : « Reannotation of genome microbial CDS ... »

« Re-annotation of genome microbial coding sequences : finding new genes and inaccurrately annotated genes », S. Bocs, A. Danchin, C. Médigue (2002) BMC Bioinformatics, 3, 5.

Le processus de réannotation semi-automatique décrit dans le chapitre précédent (voir p. 275) a été appliqué à 26 génomes procaryotes. Des informations telles que le nom de la souche, le centre de séquençage, la date de publication du génome complet, et quelques propriétés biologiques (forme, motilité, mode de vie du micro-organisme) sont synthétisées dans le tableau (Tab. 10.1 A p. 290). Les 26 résultats de comparaison du jeu de CDS annotées dans les banques (*Original Annotation* (OA)) et du jeu de CDS prédites par *AMIGene* (*AMIGene Prediction* (AP)) sont présentés dans le tableau 10.1 B p. 290 et, pour plus de détails, dans le tableau de l'annexe J p. 427.

La proportion de CDS communes (*Common Gene* (CG)) est relativement importante ; ce qui valide, en partie, le processus de réannotation. Le statut de réannotation des *annotated Gene Not Found* (GNF = OA - CG) peut prendre quatre valeurs : '*uniqBank*', '*noStatusBank*', '*suspiciousBank*' ou '*wrongBank*' (voir p. 279). L'analyse du tableau 10.1 B p. 290 révèle que malgré les énormes efforts des annotateurs des projets de séquençage des génomes, un nombre notable de gènes annotés dans ce contexte sont vraisemblablement inexacts voir complètement « faux ». Le statut de

Common name	Type var strain	Species code	Division	Doubling time	Shape	Gram	Motile	primary habitat	Human	Extremoϕ	Growth Θ	Growth pH	Pathogen Obli	Pathogen intraC	oxygen	energy	Pub year	Notable Features
Aeropyrum pernix	K1	AERPE	Crenarchaeota	200 min	C	-	Yes	marine hydroΘ vent	No	hyperΘΦ	92	7 F	F	No	aerobic	organic	1999	aerobic hyperΘΦ crenarchaeon
Agrobacterium tumefaciens	C58 (Cereon)	AGRT5	Proteobacteria (alpha)		R	-	Yes	rhizosphere, soil	No	No	26,5	P	No	No	aerobic	organic	2001	plant disease crown gall (1 linear, 1 circular chromosome)
Aquifex aeolicus	VF5	AQUEA	Aquificales		R	-	Yes	terrestrial hot springs	No	hyperΘΦ	85	F	F	No	microaero	inorganic	1998	hyper basal eubacterium
Archaeoglobus fulgidus	DSM 4304 VC-16	ARCFU	Euryarchaeota	3,5-5 h	C	-	Yes	marine hydroΘ vent	No	hyperΘΦ	83	6 F	F	No	anaerobic	inorganic	1997	hyperΘΦ, sulphate-reducing
Bacillus halodurans	C-125	BACHD	Firmicutes	2 h ?	R	+	Yes	soil	No	alkaliΦ	35	9,5 F	F	No	aerobic	organic	2000	commerically important enzyme source
Bacillus subtilis subsp. subtilis	str. 168	BACSU		26 min						No	40	5,5-8,5 OP	No				1997	saprophytic model organism
Borrelia burgdorferi	B31	BORBU	Spirochaetales	12-25 h	S	-	Yes	mammals; ticks	Yes	No	32	?	Yes	No	microaero	organic	1997	Lyme Disease
Campylobacter jejuni subsp. jejuni	NCTC 11168	CAMJE	Proteobacteria (epsilon)	colonies after 3-4 d	HR	-	Yes	animals	Yes	No	40	P	Yes	No	microaero	organic	2000	leading cause of food-poisoning
Chlamydia trachomatis	D/UW-3/CX	CHLTR	Chlamydiales	NC, cell lysis 24-30 hr	C		no	humans	Yes	No	37	P	Yes	Yes		organic	1998	genital tract infections, basal eubacteria
Chlamydophila pneumoniae	J138	CHLPN															2000	pneumonia and bronchitis
Escherichia coli	K12-MG1655	ECOLI	Proteobacteria (gamma)	21 min	R	-	Yes	animals	Yes	No	37	7,7-7,8 C	Yes	No	facultatively anaerobic	organic	1997	model organism, gut commensal inhabitant
Escherichia coli	O157:H7 EDL933	ECO57		25 min								P					2001	enterohemorrhagic pathogenic strain
Haemophilus influenzae	Rd KW20	HAEIN	Proteobacteria (gamma)	26 min	R	-	no	primates	Yes	No	36	P	Yes	No	facultatively anaerobic	organic	1995	meningitus and other human diseases
Helicobacter pylori	26695	HELPY	Proteobacteria (epsilon)	1-5 h	HR	-	Yes	humans	Yes	No	37	>2 P	Yes	No	microaero	organic	1997	stomach ulcers (acidic environment)
Helicobacter pylori	J99	HELPJ															1999	
Methanocaldococcus jannaschii	DSM 2661	METJA	Euryarchaeota	< 1 hr	C		Yes	marine hydroΘ vent	No	hyperΘΦ	85	6,5 F	F	No	anaerobic	inorganic	1996	autotrophic, archaeon
Methanothermobacter thermautotrophicus	Delta H	METTH	Euryarchaeota	8 hr	R	+	no	sewage, manure, groundwater	No	ΘΦ	68	7,2-7,6 F	F	No	anaerobic	inorganic organic	1997	isolated from sewage sludge, lithoautotrophic (Methanobacterium thermoautotrophicum)
Mycobacterium tuberculosis	CDC1551	MYCTC	Firmicutes	14-15 hr	R	+	no	humans	Yes	No	37	6,4-7 P	Yes	No	aerobic	organic	2001	Tuberculosis, clinical strain
Mycobacterium tuberculosis	H37Rv	MYCTU															1998	Tuberculosis
Mycoplasma genitalium	G-37	MYCGE	Firmicutes	1-6 hr***	FL	-	Yes	animals	Yes	No	37	P	Yes	No	facultatively anaerobic	organic	1995	model for a minimal cell
Mycoplasma pneumoniae	M129	MYCPN						humans									1996	ATCC 29342
Neisseria meningitidis	B MC58	NEIMA	Proteobacteria (beta)	30 min	C	-	no	humans	Yes	No	36	7,2 P	Yes	No	aerobic	organic	2000	cause of meningitidis
Neisseria meningitidis	A Z2491	NEIMB																
Photorhabdus luminescens subsp. laumondii	TTO1	PHOLU	Proteobacteria (gamma)	40-45 min	R	-	Yes	Nematodes	No	No	32	7 ? MP	No	No	facultatively anaerobic	organic	2003	symbiont of nematodes entomopathogen bioluminescent
Pyrococcus abyssi	GE5	PYRAB	Euryarchaeota	33 min	C	-	Yes	marine hydroΘ vent	No	hyperΘΦ	96	9 F	F	No	anaerobic	organic	1999	3500 m marine hot-spring
Pyrococcus horikoshii	OT3	PYRHO									98	7					1998	hyperthomophile
Rickettsia prowazekii	Madrid E	RICPR	Proteobacteria (alpha)	NC, 9 hr in chicken embryo cell	R	-	no	humans	Yes	No	35	P	Yes	Yes	aerobic	organic	1998	endocellular parasite, mitochondria sister taxa, non-coding DNA high proportion
Salmonella enterica subsp. enterica	Typhi CT18	SALTI	Proteobacteria (gamma)	32 min	R	-	Yes	humans	Yes	No	37	7-7,5 P	Yes	No	facultatively anaerobic	organic	2001	typhoid fever
Salmonella enterica subsp. enterica	Typhimurium LT2	SALTY						animals										leading cause of human gastroenteritis, mouse model of human typhoid fever, SGSC1412
Sulfolobus solfataricus	P2	SULSO	Crenarchaeota		C	-	Yes?	solfatara field	No	hyperΘΦ	85	4,5 F	F	No	aerobic	inorganic	2001	model crenarchaeon
Synechocystis sp.	PCC 6803	SYNY3	Cyanobacteria	~ 1 hr	R	-	Yes	freshwater	No	No	26	>6 F	F	No	aerobic	photosynthetic	1996	photosynthetic bacterium
Thermotoga maritima	MSB8	THEMA	Thermotogales	75 min	R	-	Yes	marine hydroΘ vent	No	hyperΘΦ	80	7 F	F	No	anaerobic	organic	1999	basal thermophile eubacteria
Treponema pallidum		TREPA	Spirochaetales	NC	S	-	Yes	humans	Yes	No	34	P	Yes	No	microaero	organic	1998	Syphilis
Ureaplasma urealyticum	3 1	UREUR	Firmicutes	1-6 hr***	C	-	no	humans	Yes	No	37	5 OP	Yes	No	microaero	organic	2000	mucosal pathogen of humans
Vibrio cholerae	El Tor N16961	VIBCH	Proteobacteria (gamma)	3-4 hr	R	-	Yes	aquatic	No	No	25	6-10 PF	No	No	facultatively anaerobic	organic	2000	cause of cholera, small chromosome = megaplasmid?
Yersinia pestis	Orientalis CO92	YERPE	Proteobacteria (gamma)	1,5-2 hr	R	-	no	animals	Yes	No	28,5	7,2-7,4 P	No ?	No?	facultatively anaerobic	organic	2001	plague, primarily a rodent pathogen, transmitted by fleas
Yersinia pseudotuberculosis	IP32953	YERPS		1 hr			Yes						No				2004	Soil and water-borne enteropathogen responsible for mesenteric adenite

FIG. 10.1 – A) Description de génomes procaryotes

Ces données ont été extraites du site GenomeMine (`http://www.genomics.ceh.ac.uk/cgi-bin/gmine/gminemenu.cgi`). les bactéries sont en rose et les archées en vert. Quand l'information n'est pas accessible les cases sont vides. *Doubling time in culture* : *not cultivable* (NC). *Shape* : *cocci* (C), *filamentous* (Fi), *flask-shaped* (FL), *helical rods* (HR), *prosthecate* (P), *rods* (R), *spiral* (S). *Primary habitat* : *Marine hydrothermal vents*. *Extremophiles* : *This category includes thermophiles (organisms whose optimal temperature is >40C) as well as extremophiles (based on Madigan 2000), Hyperthermophilic (>80C), Thermophilic (>48C), Alkaliphilic (>8.5), Acidophilic (<4), Halophilic (>15%).*

Species code	Length (Mb)	G+C (%)	Number of CDSs OA	Number of CDSs AP	Number of CDSs CG (%)	Annotated GNF (% vs OA) Total (OA -CC)	Pc <0.2	Status WRONG	Status SUSPI	Pc >=0.4	Potential NG (% vs AP) Total (AP-CC)	Pc >=0.4	Status NEW	Status AMBIG
AERPE	1.67	56	2694	1721	1545 (57.3)	42.65	**42.24**	33.96	4.94	0.00	10.23	**8.08**	2.56	5.35
AQUAE	1.55	43	1522	1713	1511 (99.3)	0.72	**0.72**	0.13	0.13	0.00	11.79	**10.57**	5.02	5.55
ARCFU	2.18	49	2436	2459	2360 (96.9)	3.12	**2.67**	0.78	0.94	0.04	4.03	**2.11**	0.85	1.22
BORBU	0.91	29	851	830	797 (93.7)	6.35	**5.99**	3.06	1.53	0.00	3.98	**2.53**	0.48	1.81
CAMJE	1.64	31	1647	1620	1617 (98.2)	1.82	**1.52**	0.55	0.49	0.00	0.19	**0.19**	0.00	0.19
CHLPN	1.23	41	1074	1065	1024 (95.3)	4.66	**4.56**	0.09	0.00	0.00	3.85	**2.25**	0.56	1.60
CHLTR	1.04	41	893	909	870 (97.4)	2.58	**2.24**	0.34	0.56	0.00	4.29	**2.09**	0.77	1.21
ECOLI	4.63	51	4289	4100	3959 (92.3)	7.69	**7.48**	1.42	2.12	0.00	3.44	**1.80**	0.73	1.02
HAEIN	1.83	38	1737	1765	1721 (99.1)	0.92	**0.75**	0.40	0.17	0.00	2.49	**1.30**	0.51	0.79
HELPJ	1.64	39	1482	1493	1447 (97.6)	2.36	**2.09**	0.13	0.20	0.00	3.08	**2.08**	0.40	1.54
HELPY	1.66	39	1588	1567	1514 (95.3)	4.66	**4.28**	1.89	0.44	0.00	3.38	**1.91**	0.70	1.08
METJA	1.66	31	1723	1766	1705 (99.0)	1.04	**0.99**	0.12	0.70	0.00	3.45	**2.38**	0.85	1.53
METTH	1.75	50	1869	1841	1793 (95.9)	4.07	**4.01**	1.82	1.02	0.00	2.61	**1.36**	0.16	1.20
MYCGE	0.58	32	483	550	474 (98.1)	1.86	**1.86**	0.41	0.00	0.00	13.82	**8.55**	6.00	2.18
MYCPN	0.81	40	688	805	664 (96.5)	3.49	**3.34**	0.29	0.87	0.00	17.52	**11.80**	4.60	6.96
MYCTU	4.41	66	3913	4096	3746 (95.7)	4.27	**3.71**	0.64	1.61	0.08	8.54	**3.83**	1.32	2.34
NEIMA	2.18	52	2063	1908	1802 (87.3)	12.65	**11.63**	3.64	1.65	0.00	5.56	**2.10**	0.63	1.31
NEIMB	2.27	52	2128	1960	1810 (85.1)	14.94	**14.47**	5.83	2.16	0.00	7.65	**4.39**	2.55	1.63
PYRAB	1.76	45	1764	1856	1706 (96.7)	3.29	**3.29**	0.00	0.40	0.00	8.08	**4.85**	1.45	3.39
PYRHO	1.74	42	2059	1813	1643 (79.8)	20.20	**19.86**	14.67	1.31	0.00	9.38	**5.90**	2.76	3.09
RICPR	1.10	29	834	886	818 (98.1)	1.92	**1.68**	0.72	0.60	0.00	7.67	**5.76**	2.37	3.39
SYNY3	3.57	48	3163	3111	2965 (93.7)	6.26	**6.01**	0.03	1.20	0.00	4.69	**2.06**	0.51	1.51
THEMA	1.86	46	1872	1876	1816 (97.0)	2.99	**2.78**	1.12	0.85	0.00	3.20	**1.39**	0.48	0.85
TREPA	1.14	53	1040	1034	964 (92.7)	7.31	**6.92**	3.56	2.40	0.00	6.77	**3.00**	0.68	2.03
UREPA	0.75	25	612	608	589 (96.2)	3.76	**3.59**	0.49	1.14	0.00	3.13	**1.97**	0.00	1.97
VIBCH	4.03	47	3882	3857	3568 (91.9)	8.09	**5.51**	2.32	2.01	0.00	7.49	**3.37**	0.21	3.09

FIG. 10.1 – B) Comparison of the microbial genes annotated in GenBank files with the CDS predicted by the *AMIGene* strategy (*Article III* Additional File 1 p. 289)

Interaction : this is the primary ecological type that the organism is involved in. In rare cases, there are two interactions listed, for instance for Vibrio, which has different interactions for different parts of its life cycle. Opportunistic pathogens is not noted here, as this is not the major interaction. An opportunistic pathogen is one that is not listed as "P" but does cause disease (a "yes" in the disease column). This is recorded by strain, as some strains of the same species differ in whether they are commensals or pathogens. Pathogen (P), predator (D), free-living (F), mutualist (both host and bacterium benefit, M), commensal (probably does not affect the host, C). Obligate : If the organism is listed as a P, M, or C, then does it require a host for growth ? (As far as people seem to know ; often the organisms have not been looked for in other places.). The "free-living" is for all "F" in interaction types, except for Bdellobvibrio, which is a predator ("D") but also free-living.

réannotation des *potential New Gene* (NG = AP - CG) peut prendre quatre valeurs : '*uniqAGC*', '*noStatusAGC*', '*ambiguousAGC*' ou '*newAGC*' (voir p. 279). L'analyse des nouveaux gènes prédits, montre comme attendu, que de nombreux petits gènes ont échappé à l'annotation. Dans la plupart des cas, ces nouveaux gènes révèlent des décalages du cadre de lecture (erronés ou authentiques). Des nouveaux gènes complètement inattendus ont aussi été identifiés.

Ce travail permet donc d'avoir une vision plus juste des génomes procaryotes complets. Les NG qui ont été intégrés dans Swiss-Prot valide aussi le processus de réannotation. A l'exception de quelques rares cas, les différences entre nos prédictions et les annotations sont dues à :

1. la nature du processus d'annotation (hétérogénéité d'annotation),
2. des propriétés particulières de certains génomes (hétérogénéité de composition des CDS, plasticité, pourcentage en G+C du chromosome).

Ainsi, il est absolument nécessaire d'optimiser la disponibilité et la qualité des données (*i.e.* valoriser le travail des experts pour une meilleure exploitation des annotations). D'une part, des coopérations étroites entre les scientifiques peuvent permettre de résoudre les problèmes de compatibilité syntaxiques et sémantiques des annotations. D'autre part, la mise à jour et le nettoyage régulier des annotations des banques peuvent permettre de diminuer le nombre d'erreurs et d'éviter leur propagation.

10.2 Nouveaux résultats dans PkGDB

GO type	Reannot status	Number
CDS fCDS	common	182783
	new	1606
	ambiguous	936
	noStatusAGC	354
	uniqAGC	12085
	wrong	785
	suspicious	3233
	noStatusBank	2271
	uniqBank	1622
rRNA		827
tRNA		2733

TAB. 10.1 – Statuts de 58 génomes procaryotes réannotés dans PkGDB au 1er janvier 2005

La réannotation des 26 génomes procaryotes nous a permis de mettre en lumière quelques lacunes concernant les méthodes et la gestion des données. Ainsi un certain nombre de modifications ont été apportées au processus de réannotation afin d'améliorer la qualité des annotations :

- Corriger manuellement, si possible, les bornes des CDS des banques qui ne correspondent pas à notre définition de CDS en raison d'une erreur d'annotation ou d'un décalage du cadre de lecture (*e.g.* une CDS qui ne finit pas par un codon de terminaison, ou qui en contient plusieurs, ou dont la longueur n'est pas multiple de trois). Ce travail de corrections manuelles est facilité par l'utilisation d'une interface graphique qui permet de visualiser et de modifier les bornes des CDS non conventionnelles des banques contenues dans PkGDB.
- Prendre en compte l'hétérogénéité des CDS du point de vue de leur usage des codons synonymes (*AMIMat*).
- Améliorer les heuristiques d'*AMIGene* (*e.g.* ne pas éliminer les chevauchements entre CDS '*sure*' et '*probable*' dans la même orientation).
- Valider statistiquement le paramétrage d'*AMIGene* (utilisation des quatre codons d'initiation [ACGT]TG) et ajuster empiriquement les paramètres de la méthode *SWAN* (*e.g.* $sure\_Pc = 0{,}7$ au lieu de 0,4 et $prob\_Pc = 0{,}4$ au lieu de 0,2).
- Stocker dans PkGDB, non seulement les annotations des banques, mais aussi les prédictions *AMIMat* et les statuts de réannotation. Autrement dit, intégrer dans un même pipeline qui interagit avec PkGDB, sans en être pour autant dépendant (modularité) : l'analyse des fichiers *RefSeq*, la correction manuelle des annotations, la construction des matrices de transition (*AMIMat*), la prédiction de CDS (*AMIGene*), la comparaison des annotation (*CompAnnot*), la recherche de similitudes dans les banques de séquences protéiques pour les polypeptides correspondant aux CDS trouvées de façon unique dans les annotations des banques ou dans les prédictions *AMIGene* (Blast2P de la plate-forme Biofacet contre la SWALL), l'attribution des statuts à certaines de ces CDS (*SWAN*), le chargement de tous les objets génomiques, les

analyses ultérieures de ces objets pour finalement visualiser une représentation synthétique de tous ces résultats *via* le serveur *MaGe*.

Au 1er janvier 2005, 58 génomes procaryotes ont été réannoté avec cette nouvelle version du processus, ce qui représente 182783 CDS '*common*', 7911 CDS uniques aux banques, 14981 CDS uniques à *AMIGene*, 2733 tRNA et 827 rRNA dans la table [Genomic_Object] de la base PkGDB (Fig. 10.1 p. 293). A ces données s'ajoutent d'autres objets génomiques des banques (*e.g.* RBS, misc_RNA, sc_RNA, misc_feature, repeat_unit) et des résultats d'autres méthodes de prédiction (*e.g.* prédiction de décalage du cadre de lecture, recherche de répétitions, de similitudes contre la SWALL, de motifs protéiques, d'orthologues, de paralogues, de synténies, d'activité enzymatique).

Le tableau 10.2 p. 295 permet de comparer des résultats de la nouvelle version du processus de réannotation avec ceux de l'ancienne version (les lignes où le code d'espèce est suivi d'une astérix dans le tableau 10.2 p. 295 correspondent aux anciens résultats du tableau 10.1 B p. 290). Pour tous les génomes, nous avons maintenant plus de 92% de CDS communes aux annotations des banques et aux prédictions *AMIGene*. La progression des résultats la plus spectaculaire est celle associée aux génomes des *Neisseria* (voir p. 311). Le cas des archaea mériterait une étude plus approfondie, en particulier, de l'usage des codons synonymes des CDS [Ochman *et al.*, 2000]. En effet, les génomes d'archaea sont ceux dont les CDS montrent en moyenne la plus grande variation de pourcentage en G+C (déviation standard moyenne de 4,13 ; Tab. 10.3 p. 297 [Hsiao *et al.*, 2003]). Il existe bien chez les archaea des régions atypiques plus difficiles à prédire que les régions classiques (Fig. 10.2 p. 296). Cependant, nous utilisons actuellement, une seule matrice pour prédire les CDS des génomes d'archaea avec un jeu de paramètres[1] indépendant du contenu en G+C. Les résultats de réannotation se sont améliorés dans le cas de *M. jannaschii* et *P. abyssi* mais pas dans le cas de *Methanothermobacter thermoautotrophicus*. Les génomes de *M. jannaschii* et *P. abyssi* sont ceux pour lesquels nous obtenons la plus grande proportion de CDS communes (avec une progression resp. de 98,96 à 99,20 et de 96,71 à 99,32%). Les amélioration d'*AMIGene* et les valeurs de paramètre choisies suite à la validation automatique des paramètres d'*AMIGene* sont vraisemblablement à l'origine de cette progression. Pour *M. thermoautotrophicum*, le pourcentage de CDS communes chute de 95,93 à 95,30, certainement parce que, dans ces conditions, moins de CDS sont prédites par *AMIGene*.

						Number of CDS						Annotated GNF (% vs OA)												Potential NG (% vs AP)									
Species	Strain	S_id	Length	Cl	G+C	Bank			AP	CC		OA-CC		low Pc		Status						high Pc		AP-CC		high Pc		Status					
code			(pb)	Nb	(%)	total	FS	OA			(%)		(%)		(%)	W	(%)	S	(%)	NS	(%)		(%)		(%)	high	(%)	N	(%)	A	(%)	NS	(%)
AGRT5	C58 Cereon	28	2841581	2	59,40	2721	0	2721	2902	2649	97,35	72	2,65	55	2,02	9	0,33	43	1,58	3	0,11	0	0,00	253	8,72	34	1,17	26	0,90	7	0,24	1	0,03
BACHD	C-125	1	4202353	3	43,69	4067	8	4056	4470	3998	98,57	58	1,43	43	1,06	12	0,30	24	0,59	7	0,17	0	0,00	472	10,56	42	0,94	26	0,58	16	0,36	0	0,00
BACSU*	168		4214814	1	43,51			4105	4193	3956	96,37	149	3,63	123	3,00	29	0,71	26	0,63	68	1,66	1	0,02	237	5,65	120	2,86	50	1,19	66	1,57	4	0,10
BACSU	168	9	4214630	3	43,51	4107	5	4107	4586	4066	99,00	41	1,00	41	1,00	10	0,24	15	0,37	15	0,37	0	0,00	520	11,34	83	1,81	47	1,02	35	0,76	1	0,02
BORBU*	B31		910724	1	28,60			851	830	797	93,65	54	6,35	51	5,99	26	3,06	13	1,53	12	1,41	0	0,00	33	3,98	21	2,53	4	0,48	15	1,81	2	0,24
BORBU	B31	26	910724	2	28,60	852	0	852	854	821	96,36	31	3,64	25	2,93	4	0,47	18	2,11	3	0,35	0	0,00	33	3,86	14	1,64	4	0,47	10	1,17	0	0,00
CHLTR*	NULL		1042519	1	41,30			893	909	870	97,42	23	2,58	20	2,24	3	0,34	5	0,56	12	1,34	0	0,00	39	4,29	19	2,09	7	0,77	11	1,21	1	0,11
CHLTR	NULL	18	1042519	2	41,30	896	3	895	945	882	98,55	13	1,45	11	1,23	1	0,11	5	0,56	5	0,56	0	0,00	63	6,67	9	0,95	3	0,32	6	0,63	0	0,00
CHLPN*	J138		1226565	1	40,60			1074	1065	1024	95,34	50	4,66	49	4,56	1	0,09	0	0,00	48	4,47	0	0,00	41	3,85	24	2,25	6	0,56	17	1,60	1	0,09
CHLPN	J138	19	1226565	2	40,60	1069	13	1069	1129	1045	97,75	24	2,25	21	1,96	0	0,00	12	1,12	9	0,84	0	0,00	84	7,44	13	1,15	9	0,80	4	0,35	0	0,00
ECOLI*	K-12		4639221	1	50,79			4289	4100	3959	92,31	330	7,69	321	7,48	61	1,42	91	2,12	169	3,94	0	0,00	141	3,44	74	1,80	30	0,73	42	1,02	2	0,05
ECOLI	K-12	13	4639221	3	50,79	4273	258	4264	4553	4201	98,52	63	1,48	44	1,03	3	0,07	4	0,09	37	0,87	0	0,00	352	7,73	53	1,16	47	1,03	5	0,11	1	0,02
ECO57	EDL933	2	5528445	3	50,38	5351	61	5339	5768	5181	97,04	158	2,96	119	2,23	14	0,26	39	0,73	66	1,24	1	0,02	587	10,18	55	0,95	42	0,73	12	0,21	1	0,02
METJA*	DSM 2661		1664970	1	31,40			1723	1766	1705	98,96	18	1,04	17	0,99	2	0,12	12	0,70	3	0,17	0	0,00	61	3,45	42	2,38	15	0,85	27	1,53	0	0,00
METJA	DSM 2661	16	1664970	1	31,40	1741	25	1741	1786	1727	99,20	14	0,80	9	0,52	1	0,06	7	0,40	1	0,06	0	0,00	59	3,30	17	0,95	6	0,34	11	0,62	0	0,00
METTH*	Delta H		1751377	1	49,50			1869	1841	1793	95,93	76	4,07	75	4,01	34	1,82	19	1,02	22	1,18	0	0,00	48	2,61	25	1,36	3	0,16	22	1,20	0	0,00
METTH	Delta H	15	1751377	1	49,50	1873	3	1873	1827	1785	95,30	88	4,70	84	4,48	6	0,32	42	2,24	36	1,92	0	0,00	42	2,30	15	0,82	5	0,27	9	0,49	1	0,05
MYCTU	CDC1551	6	4403836	4	65,61	4279	88	4273	4736	4002	93,66	271	6,34	223	5,22	28	0,66	143	3,35	52	1,22	2	0,05	734	15,50	65	1,37	41	0,87	20	0,42	4	0,08
MYCTU**	CDC1551	6	4403836	4	65,61	4279	88	4273	4448	4035	94,43	238	5,57	208	4,87	38	0,89	118	2,76	52	1,22	2	0,05	413	9,29	89	2,00	54	1,21	31	0,70	4	0,09
MYCTU***	CDC1551	6	4403836	4	65,61	4279	88	4273	4448	4035	94,43	238	5,57	208	4,87	45	1,05	125	2,93	38	0,89	2	0,05	413	9,29	89	2,00	59	1,33	26	0,58	4	0,09
MYCTU*	H37Rv		4411529	1	65,61			3913	4096	3746	95,73	167	4,27	145	3,71	25	0,64	63	1,61	57	1,46	3	0,08	350	8,54	157	3,83	54	1,32	96	2,34	7	0,17
MYCTU	H37Rv	7	4411532	4	65,61	3996	27	3996	4739	3946	98,75	50	1,25	41	1,03	1	0,03	28	0,70	12	0,30	1	0,03	793	16,73	28	0,59	16	0,34	10	0,21	2	0,04
MYCTU**	H37Rv	7	4411532	4	65,61	3996	27	3996	4387	3959	99,07	37	0,93	31	0,78	0	0,00	6	0,15	25	0,63	1	0,03	428	9,76	41	0,93	19	0,43	20	0,46	2	0,05
MYCTU***	H37Rv	7	4411532	4	65,61	3996	27	3996	4387	3959	99,07	37	0,93	31	0,78	0	0,00	6	0,15	25	0,63	1	0,03	428	9,76	41	0,93	15	0,34	24	0,55	2	0,05
NEIMA*	Z2491		2184406	1	51,80			2063	1908	1802	87,35	261	12,65	240	11,63	75	3,64	34	1,65	131	6,35	0	0,00	106	5,56	40	2,10	12	0,63	25	1,31	3	0,16
NEIMA	Z2491	24	2184406	3	51,80	2155	125	2147	2284	2069	96,37	78	3,63	59	2,75	6	0,28	44	2,05	9	0,42	1	0,05	215	9,41	58	2,54	16	0,70	39	1,71	3	0,13
NEIMB*	MC58		2272351	1	51,53			2128	1960	1810	85,06	318	14,94	308	14,47	124	5,83	46	2,16	138	6,48	0	0,00	150	7,65	86	4,39	50	2,55	32	1,63	4	0,20
NEIMB	MC58	34	2272351	3	51,53	2224	158	2211	2383	2053	92,85	158	7,15	144	6,51	41	1,85	82	3,71	21	0,95	1	0,05	330	13,85	116	4,87	57	2,39	57	2,39	2	0,08
PHOLU	TT01	10	5688987	3	42,83	4905	222	4905	5883	4839	98,65	66	1,35	34	0,69	0	0,00	24	0,49	10	0,20	1	0,02	1044	17,75	70	1,19	17	0,29	51	0,87	2	0,03
PSEHA	TAC 125	30	3186259	3	40,10				3068																								
PSEHA	TAC 125	31	634391	3	39,37				565																								
PYRAB*	NULL		1765118	1	44,70			1764	1856	1706	96,71	58	3,29	58	3,29	0	0,00	7	0,40	51	2,89	0	0,00	150	8,08	90	4,85	27	1,45	63	3,39	0	0,00
PYRAB	NULL	17	1765118	1	44,70	1770	3	1768	1908	1756	99,32	12	0,68	11	0,62	1	0,06	4	0,23	6	0,34	1	0,06	152	7,97	75	3,93	60	3,14	15	0,79	0	0,00
SALTY*	LT2		4857432	3	52,22			4473	4689	4366	97,61	107	2,39	96	2,15	9	0,20	12	0,27	75	1,68	0	0,00	323	6,89	55	1,17	29	0,62	26	0,55	0	0,00
SALTY	LT2	35	4857432	3	52,22	4516	69	4507	4704	4389	97,38	118	2,62	94	2,09	7	0,16	23	0,51	64	1,42	0	0,00	315	6,70	34	0,72	25	0,53	9	0,19	0	0,00
SULSO	P2	14	2992245	1	35,80	2978	28	2977	3322	2926	98,29	51	1,71	44	1,48	1	0,03	15	0,50	28	0,94	0	0,00	396	11,92	131	3,94	90	2,71	39	1,17	2	0,06
YERPE*	CO92		4653728	1	47,64			4009	4603	3910	97,53	99	2,47	92	2,29	55	1,37	15	0,37	22	0,55	0	0,00	693	15,06	228	4,95	32	0,70	195	4,24	1	0,02
YERPE	CO92	3	4653728	3	47,64	4110	355	4108	4336	3981	96,91	127	3,09	108	2,63	2	0,05	70	1,70	36	0,88	0	0,00	355	8,19	30	0,69	13	0,30	14	0,32	3	0,07
YERPS	IP32953	11	4744671	3	47,61	3978	37	3976	4282	3916	98,49	60	1,51	48	1,21	1	0,03	27	0,68	19	0,48	0	0,00	366	8,55	37	0,86	14	0,33	22	0,51	1	0,02

TAB. 10.2 – PkGDB and comparison of the microbial genes annotated in GenBank files with the CDS predicted by the *AMIGene* strategy

Abbreviations : CDS, coding sequence, fCDS coding fragment ; Nb Cl, cluster number ; OA, original annotation ; AP, *AMIGene* prediction ; CG, CDS common to both OA and AP ; GNF, gene not found ; NG, new gene ; Pc, coding average probability of a CDS ; S, suspicious reannotation status ; W, wrong ; A, ambiguous ; N, New. Les codes d'espèce soulignés correspondent aux archaea. (*) Une seule matrice 'anciens paramètres' à l'exception de SALTY qui est le premier génome à avoir été réannoté avec plusieurs matrices. (**) PkGDB : nouvelles matrices (ordre 6 et NC 333 kb) et nouveaux paramètres. (***) Mêmes critères que (**) et utilisation des filtres Seg et Xnu.

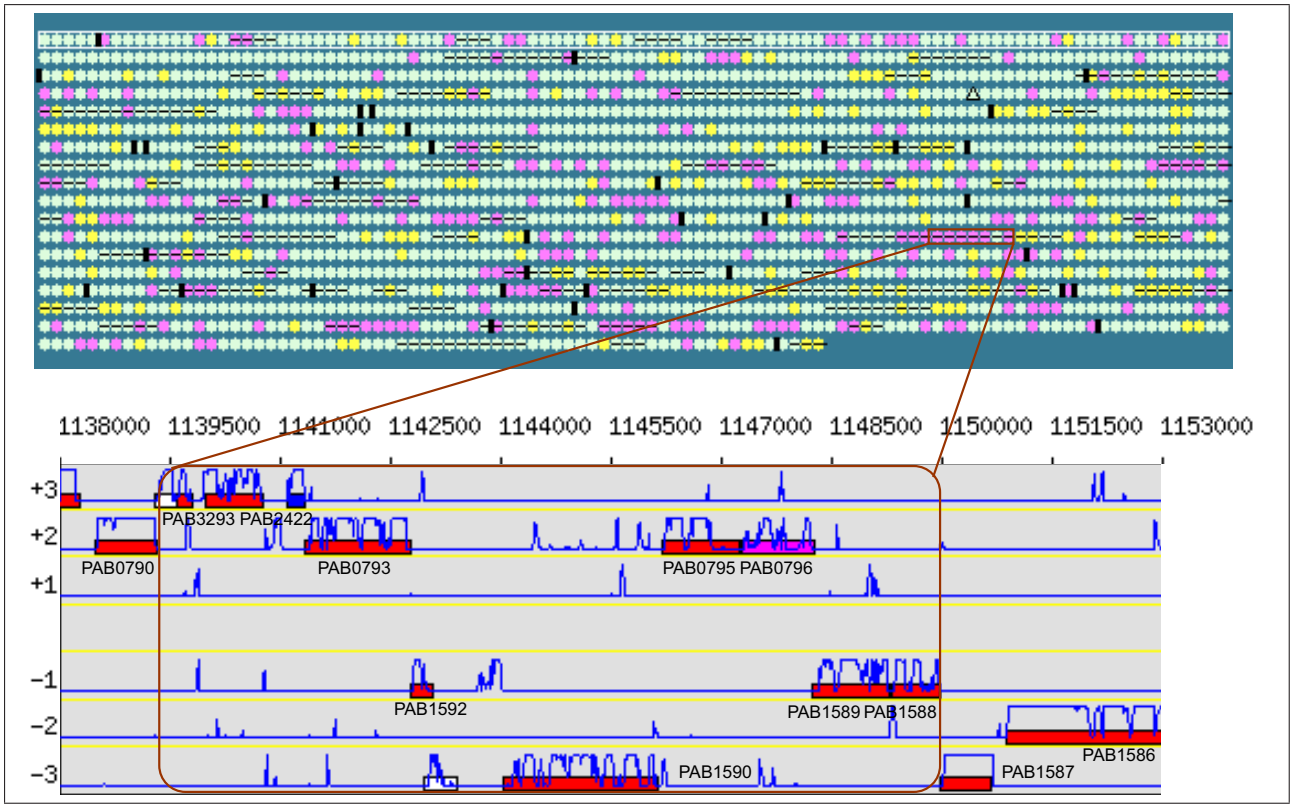


FIG. 10.2 – Région atypique chez *Pyrococcus abyssi*

La première image a été téléchargée à partir du site d'IslandPath [Hsiao *et al.*, 2003] : A yellow circle indicates its %G+C value is bigger than the high value. A green circle indicates its %G+C value is bigger than the low value and smaller than the high value. A pink circle indicates its %G+C value is smaller than the low value. A strike line across the circles indicates the region has dinucleotide bais above 1 STD DEV. A black verticle bar indicates a transfer RNA gene lies between the two ORFs. A purple verticle bar indicates a ribosomal RNA gene lies between the two ORFs. A deep blue verticle bar indicates both a tRNA and rRNA gene lie between the two ORFs. A black square indicates the dot is annotated as a transposase. A black triangle indicates the dot is annotated as an integrase.

Les CDS atypiques dans leur composition en oligonucléotides sont encadrées sur cette région de *P. abyssi*. C'est un groupe de CDS hypothétiques.

Pathogenic Bacteria	%G+C (all ORF) Mean	S.D.	%G+C (ORF >300bp) Mean	S.D.	Genome Dinucleotide Bias Mean	S.D.
Mean	**44,53**	4,07	44,98	3,69	45,69	20,53
S.D.	11,02	0,97	11,28	0,89	4,50	3,34
Minimum	25,80	2,30	25,80	2,20	34,20	12,70
Maximum	67,00	7,60	67,30	6,90	54,20	30,10

Non pathogenic Bacteria	%G+C (all ORF) Mean	S.D.	%G+C (ORF >300bp) Mean	S.D.	Genome Dinucleotide Bias Mean	S.D.
Mean	49,60	4,05	49,85	3,80	43,75	18,48
S.D.	14,06	0,77	14,15	0,66	3,39	2,74
Minimum	23,40	3,00	23,40	2,80	35,30	13,40
Maximum	72,10	6,40	72,30	5,40	52,60	25,50

Archaea	%G+C (all ORF) Mean	S.D.	%G+C (ORF >300bp) Mean	S.D.	Genome Dinucleotide Bias Mean	S.D.
Mean	46,34	4,34	46,48	4,13	46,13	19,69
S.D.	9,41	0,74	9,53	0,74	3,78	2,01
Minimum	31,70	3,40	31,70	3,10	40,00	16,10
Maximum	67,90	5,60	68,40	5,50	53,60	22,90

TAB. 10.3 – Contenu en G+C des ORF et biais en dinucléotides des génomes procaryotes
Ces résultats ont été calculés à partir des données du site d'IslandPath `http://pathogenomics.sfu.ca/islandpath/current/IPindex.pl`. Par exemple pour les bactéries pathogènes, 44,53 est la moyenne des moyennes calculées sur les ORFs pour chaque génome.

10.3 Complément d'information sur la réannotation de génomes de *Protéobactéries*

10.3.1 *E. coli* K-12

Gènes uniques aux annotations

Plus de 17% des gènes d'*E. coli* K-12 ont probablement été transférés horizontalement à partir d'autres génomes [Lawrence & Ochman, 1998]. La version d'*AMIGene* qui a servi à la réannotation des 26 génomes procaryotes complets n'utilisait pas de matrice de transition spécifique de ce groupe de gènes [Borodovsky *et al.*, 1995], ce qui peut expliquer les 7,69% de GNF (TAB. 10.1 B p. 290 et *Article III* p. 289). En effet, la majorité des GNF ont une faible probabilité moyenne de codage (7,48%) mais n'ont pas de statut (4,15%) et la majorité des GNF, qui ont un statut, sont '*suspiciousBank*' (2,12%). Ces gènes correspondent soit à des protéines hypothétiques, soit à des fonctions associées aux IS, aux fimbriae, aux lipoprotéines, à certains régulateurs ou antigènes.

La majorité des GNF '*wrongBank*' [Blattner *et al.*, 1997] ont été retirés du jeu de réannotation fourni par la base spécialisée EcoGene [Rudd, 1998]. Parmi les 61 GNF '*wrongBank*' (1,42%), 36

[1] *climb_P*=0.965, *diff_Pc*=0.07, *sure_Pc*1=0.7, *prob_L*1=150, *sure_L*=60, *prob_Pc*=0.4, *sure_Pc*2=0.7, *prob_L*2=150, *sure_ss_I*=0.1, *sure_os_I*=0.55, *sure_prob_O*=0.1, *prob_glob_IO*=0.75

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
213928	R	216	YAEP_ECOLI	66	HYPOTHETICAL 7.2 KDA PROTEIN IN ROF-CUTF INTERGENIC REGION (EG13222 B0189.1)	3.1e-29	98	b0189	hypothetical protein (yaeO)	b0190	hypothetical protein (yaeQ)	YAEP B0189.1 1 66 7 72
279251	R	342	Q57247	112	HYPOTHETICAL 12.9 KDA PROTEIN	7.9e-56	100	b0265	IS1 protein InsA (insA_2)	b0266	hypothetical protein (yagB)	Q57247 1 112 3 114
547581	D	258	YLBE_ECOLI	419	HYPOTHETICAL 45.0 KDA PROTEIN in FDRA-ARCC intergenic region (ORF2 EG13398)	3.7e-42	98	b0518	involved in protein transport; peptide secretion(fdrA)	b0519	hypothetical protein (ylbE)	YLBE_ECOLI 1 86 b0519 1 86
574836	D	174	P77184	153	from bases 1425674 to 1439104 of the complete genome (section 124 of 400) (b1371)	1.3e-21	79	b0552	IS5 transposase (trs5_2)	b0553	outer membrane porin protein; locus of qsr prophage (nmpC)	b1371 34 91 1 58
613213	D	168	YBDZ_ECOLI	72	HYPOTHETICAL 8.3 KDA PROTEIN IN FES-ENTF INTERGENIC REGION (EG14353)	2.8e-28	100	b0585	enterochelin esterase (fes)	b0586	ATP-dependent serine activating enzyme (enterobactin synthase component F)	YBDZ_ECOLI 17 72 1 56
631405	D	195	YBDD_ECOLI	65	HYPOTHETICAL 7.5 KDA PROTEIN IN CSTA-AHPC INTERGENIC REGION (EG11107)	9.9e-33	100	b0598	carbon starvation protein (cstA)	b0599	putative oxidoreductase (ybdH)	YBDD_ECOLI 1 65 1 65
1049056	D	273	AAF70015	91	IS1 TRANSPOSASE (INSA)	7.2e-46	98	b0987	hypothetical protein (ymcD)	b0988	IS1 protein InsB (insB_4)	INSA 1 91 1 91
1051290	D	171	YCCL_ECOLI	57	HYPOTHETICAL 6.6 KDA PROTEIN IN SFA-TORS INTERGENIC REGION (YCCL EG13233)	1.3e-23	98	b0991	suppresses fabA and ts growth mutation (sfa SUSPICIOUS)	b0992	hypothetical protein (yccM)	YCCL_ECOLI 1 57 1 57
1210636	D	162	IDH_ECOLI	416	ISOCITRATE DEHYDROGENASE [NADP] (EC 1.1.1.42) (IDH IDP icdA_2 b1136_2)	2.3e-23	96	b1159	restriction of DNA at 5-methylcytosine residues; at locus of e14 element (mcrA)	b1160	hypothetical protein (SUSPICIOUS ycgW)	icdA_2 b1136_2 363 416 1 54
1394067	R	174	P77184	153	from bases 1425674 to 1439104 of the complete genome (section 124 of 400) (b1371)	1.3e-21	79	b1330	hypothetical protein	b1331	IS5 transposase (trs5_4)	b1371 34 91 1 58
1427573	R	747	Y206_LAMBD	206	HYPOTHETICAL PROTEIN ORF206	1.3e-53	67	b1371	hypothetical protein	b1372	putative membrane protein	5 205 45 248
1488624	R	114	G3P3_ECOLI	332	Glyceraldehyde 3-phosphate dehydrogenase C (EC1.2.1.12) (EG12103 GAPDH-C)	6.6e-14	100	b1417	glyceraldehyde 3-phosphate dehydrogenase C, interrupted (gapC_1)	b1418	cytochrome b(561) (cybB)	G3P3_ECOLI 1 38 b1417 b1416 1 38

FIG. 10.3 – CDS '*newAGC*' chez *E. coli* K-12 (pour la légende voir TAB. 10.5 p. 304)

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
2287961	D	174	P77184	153	from bases 1425674 to 1439104 of the complete genome (section 124 of 400) (b1371)	1.3e-21	79	b2192	IS5 transposase (trs5_8)	b2193	nitrate/nitrite response regulator (sensor NarQ) (narP)	b1371 34 91 1 58 1
2355836	D	183	YFAD_ECOLI	299	HYPOTHETICAL 34.9 KDA PROTEIN IN GLPC-AIS INTERGENIC REGION (EG12323)	1.1e-17	75	b2244	hypothetical protein (yfaD)	b2245	hypothetical protein	YFAD_ECOLI 244 299 1 61
2370577	D	333	P81891	111	HYPOTHETICAL 12.1 KDA PROTEIN IN PQAB 3'REGION (ORF6)	2.1e-39	74	b2257	hypothetical protein	b2258	putative transport/ receptor protein	P81891 1 109 1 109
2525964	R	216	YPEB_ECOLI	72	HYPOTHETICAL 8.4 KDA PROTEIN IN XAPB-LIG INTERGENIC REGION (EG14366)	1.1e-33	100	b2410	putative cytochrome oxidase (yfeH)	b2411	DNA ligase (lig)	YPEB_ECOLI 1 72 1 72
3128160	R	174	P77184	153	from bases 1425674 to 1439104 of the complete genome (section 124 of 400) (b1371)	1.3e-21	79	b2981	hypothetical protein	b2982	IS5 transposase trs5_9	b1371 34 91 1 58
3476232	R	198	YHEV_ECOLI	66	HYPOTHETICAL 7.6 KDA PROTEIN IN SLYD-KEFB INTERGENIC REGION (EG14364)	9.9e-33	100	b3349	FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase) (slyD)	b3350	K+ efflux; NEM-activable K+/H+ antiporter (kefB)	YHEV_ECOLI 1 66 1 66
3650688	D	174	P77184	153	from bases 1425674 to 1439104 of the complete genome (section 124 of 400) (b1371)	1.3e-21	79	b3505	IS5 transposase (trs5_11)	b3506	outer membrane protein induced after carbon starvation (slp)	b1371 34 91 1 58
3858803	R	210	MALH_FUSMR	441	Maltose-6'-phosphate glucosidase (EC 3.2.1. 122) (6-phospho-alpha -D-glucosidase)	4.6e-23	74	b3680	putative ARAC-type regulatory protein (yidL)	b3681	probable 6-phospho-beta-glucosidase (glvG)	MALH_FUSMR 372 441 b3681 1 70
3930694	D	114	KUP_ECOLI	622	KUP SYSTEM POTASSIUM UPTAKE PROTEIN (TRKD EG11541)	1.1e-13	100	b3747	low affinity potassium transport system (kup)	b3748	D-ribose high-affinity transport system; membrane-associated protein (rbsD)	KUP_ECOLI 585 622 b3747 1 38
3992135	D	201	YIFL_ECOLI	67	HYPOTHETICAL 7.2 KDA PROTEIN IN CYAY-DAPF INTERGENIC REGION (EG12353)	1.3e-32	100	b3808	hypothetical protein (SUSPICIOUS)	b3809	diaminopimelate epimerase (dapF)	YIFL_ECOLI 1 67 1 67
4005413	D	231	RHTC_ECOLI	206	THREONINE EFFLUX PROTEIN	2.4e-29	100	b3822	ATP-dependent DNA helicase (recQ)	b3823	hypothetical protein (yigJ)	RHTC_ECOLI 17 79 b3823 1 63
4006462	R	204	RHTB_ECOLI	206	HOMOSERINE/ HOMOSERINE LACTONE EFFLUX PROTEIN (EG11469)	2.2e-28	96	b3824	hypothetical protein (yigK)	b3825	lysophospholipase L(2) (pldB)	RHTB_ECOLI 1 65 b3824 1 65
4012947	R	171	DLHH_ECOLI	258	Putative carboxymethyl enebutenolidase (EC 3. 1.1.45)(dienelactone hydrolase)(DLH EG14321)	4.1e-27	100	b3829	tetrahydropteroyl triglutamate methyltransferase (metE)	b3830	putative enzyme (ysgA)	DLHH_ECOLI 202 258 b3830 1 57
4031991	D	192	TRKH_ECOLI	483	TRK SYSTEM POTASSIUM UPTAKE PROTEIN TRKH (EG11021)	1.7e-29	100	b3849	potassium uptake, requires TrkE (trkH)	b3850	protoporphyrin oxidase (hemG)	TRKH_ECOLI 420 483 b3849 3 64
4139645	R	156	YPDD_ECOLI	831	Putative phosphoenol pyruvate phosphotransferase (EC 2.7.3.9) (enzyme I) (EG14151)	3.0e-05	42	b3948	hypothetical protein (yijI SUSPICIOUS)	b3949	PTS system, fructose-like enzyme II component (frwC)	YPDD_ECOLI 2 50 b3947 3 51
4190218	R	225	THIS_ECOLI	66	THIS PROTEIN (EG14363)	3.5e-30	100	b3991	thiamin biosynthesis, thiazole moiety (thiG)	b3992	thiamin biosynthesis, thiazole moiety (thiF)	THIS_ECOLI 1 66 10 75
4311434	D	339	Q9I027	126	HYPOTHETICAL PROTEIN PA2816	2.8e-05	35	b4091	hypothetical protein (phnQ WRONG)	b4092	phosphonate metabolism (phnP)	Q9I027 33 109 37 110
4527826	R	276	PTKB_ECOLI	94	PTS system, galacticol specific IIB component (EIIB-GAT)(EC 2.7.1.69) (EG12415)	5.3e-11	35	b4304	putative PTS system enzyme IIC component (sgcC)	b4305	putative lyase/synthase (sgcX)	PTKB_ECOLI 3 92 2 90

FIG. 10.3 – CDS '*newAGC*' chez *E. coli* K-12

ont disparu de la version EcoGene17, 21 correspondent à des gènes dont la fonction est inconnue ou hypothétique et seulement 4 sont finalement de véritables gènes. Les quatre gènes *lar*, *sra*, *dicB* et *tnaC* codent respectivement une protéine de prophage intervenant dans la restriction, une protéine associée au ribosome (induite en phase stationnaire), une protéine intervenant dans le contrôle de la division cellulaire et un peptide *leader* régulant l'expression de l'opéron tryptophanase.

Des résultats similaires sont obtenus lorsqu'on prend comme référence les annotations de *GenProtEC* (version du 5 juin 2003 ; [Riley & Labedan, 1997]). *GenProtEC* suit en général les annotations déposées par F. Blattner dans INSD, la curation étant effectuée au niveau des attributs fonctionnels. Sur les 61 GNF '*wrongBank*', 40 ont une fonction inconnue (équivalant aux 36 gènes disparus d'Ecogene) et 21 ont une fonction connue. Parmi ces 21 GNF '*wrongBank*', on retrouve les quatre gènes *lar*, *sra*, *dicB* et *tnaC*. Pour les 17 restant, on trouve deux peptides *leader* (b0079 et b3782), neuf protéines[2] de prophage, deux protéines hypothétiques conservées (b2112 et b4091) et deux protéines d'autres fonctions (b0701 et b0991).

Prédiction de nouveaux gènes potentiels

Dans le cas du génome d'*E. coli* K-12, 0,73% des NG (30 CDS) ont le statut '*newAGC*' (Tab. 10.1 B p. 290). Dans la catégorie des NG ($Pc \geq 0,4$), nous avons pris soin d'éliminer les fragments de CDS générés par la présence exceptionnelle de codon opale TGA codant la sélénocystéine (notre stratégie génère deux fragments car, généralement, TGA est un codon stop). En prenant EcoGene comme base de référence au lieu du fichier INSD, nous retrouvons environ deux tiers des nouvelles annotations dans Ecogene (18 CDS). Ce résultat valide la stratégie de réannotation. Les 30 CDS '*newAGC*' sont de petite taille (longueur comprise entre 100 et 350 pb ; Tab. 10.3 p. 298). Grâce aux résultats de la stratégie *ProFED* et/ou de la recherche de similitudes, on en déduit que 14 CDS '*newAGC*' sont complètes et 16 sont des fragments (/partial) éventuellement impliquées dans des décalages du cadre de lecture (/pseudo). On observe aussi que 23 CDS '*newAGC*' codent des polypeptides qui sont similaires à des protéines dont la fonction est caractérisée et 7 sont similaires à des protéines hypothétiques et conservées. Par exemple, la CDS '*newAGC*', b4091.1, est très similaire à la protéine hypothétique C5097 d'*E. coli* O6 et à Z5694 d'*E. coli* O157 :H7. B4091.1 est située en 3′ de *rpiB* sur le brin direct et en face de *phnQ* sur le brin inverse. Or, la CDS *phnQ*, qui est annotée comme protéine très hypothétique dans Swiss-Prot, a justement le statut '*wrongBank*' et a aussi été supprimée d'EcoGene. De plus, parmi les 23 polypeptides '*newAGC*' de fonction caractérisée, 10 ont des similitudes avec des séquences de phages ou des séquences d'insertion. Par ailleurs, nous avons trouvé, comme dans la plupart des génomes analysés, certaines protéines annotées dans la banque de séquences Swiss-Prot, bien qu'elles ne figurent pas dans les annotations INSD. Elles correspondent soit à des protéines qui avaient été caractérisées lors d'études précédentes, soit à des protéines directement annotées par le groupe d'A. Bairoch (elles portent dans ce cas la mention *NOT-ANNOTATED-CDS*). Enfin, il est intéressant de discuter le cas de la CDS '*newAGC*'

[2] B0559, b1144, b1347, b1354, b1364, b1551, b1561, b1567 et b1576.

b1372.1 sur le brin inverse, qui est incluse dans b1372 sur le brin direct. B1372 est similaire à STF_LAMBD qui est une protéine de la fibre de la queue du bactériophage lambda. La séquence protéique b1372.1 possède des régions de basse complexité compositionnelle, masquées par le filtre Seg, et des répétitions de courte périodicité, masquées par le filtre Xnu. En utilisant ces filtres, aucune similitude n'est détectable (la séquence est complètement masquée). Sans ces filtres, nous trouvions une E-value de 1,3e-53 avec Y206_LAMBD et une E-value de 1,2e-6 avec STF_LAMBD. Ainsi, les séquences protéiques directe (b1372) et inverse (b1372.1) sont similaires à la même protéine STF_LAMBD (la séquence directe est similaire à la séquence complémentaire inverse). Ceci n'est possible que grâce à des séquences palindromiques du type GATC (l'inverse complémentaire est aussi GATC). Ici, nous n'avons donc plus à faire à un artefact de CDS de type ombre du codant. Cependant la forte similitude de b1372.1 avec Y206_LAMBD autorise l'hypothèse selon laquelle, chez les phages, il est possible d'avoir des cas avérés d'inclusion de CDS en sens contraire (une CDS sur le brin direct, par exemple, inclut une CDS sur le brin inverse). Depuis, l'entrée Y206_LAMBD a été supprimée de Swiss-Prot car cette CDS est en fait une « fausse » prédiction. Finalement, b1372.1 n'est pas une nouvelle CDS mais un artefact dû aux répétitions. Cependant, cet artefact a aussi été annoté chez *E. coli* O157:H7 EDL933, c'est la CDS Z1919 (Q8X4J5). La propagation des erreurs d'annotation nécessite donc un travail de curation régulier. Aussi, il semble important d'utiliser des filtres pour masquer les séquences de faible complexité compositionnelle ou contenant des régions répétées de courtes périodicité afin d'éliminer des similitudes statistiquement significatives mais biologiquement inintéressantes.

Dans le tableau 10.2 p. 295, les résultats de réannotation de la ligne ECOLI* correspondent aux annotations GenBank [Blattner *et al.*, 1997] tandis que ceux de la ligne ECOLI correspondent aux annotations d'EcoGene17 (voir p. 213 [Rudd, 2000]).

10.3.2 *S. enterica* serovar Typhimurium LT2

Les génomes des entérobactéries pathogènes *S. enterica* serovar Typhimurium LT2 et *S. enterica* serovar Typhi CT18 ont été respectivement séquencés au Genome Sequencing Center (GSC) de l'Université de Washington [McClelland *et al.*, 2001] et au Sanger Center (UK) [Parkhill *et al.*, 2001a]. En 2002, nous avons entamé une collaboration avec M. McClelland du Sidney Kimmel Cancer Center (Californie), pour la réannotation de ces deux génomes (Tab. 10.4 p. 302 [Jackson, 2002]).

Le génome de *S. enterica* serovar Typhimurium LT2 est le premier à avoir été réannoté avec des matrices de transition spécifiques de la classe d'usage des codons synonymes (trois pour les génomes d'entérobactéries comme *E. coli* K-12). Ainsi, on comprend qu'il y ait peu de différences entre les résultats SALTY* et SALTY du tableau 10.2 p. 295 puisque, dans les deux cas, nous avions trois matrices de transition ; les seuls changements dans le processus d'annotation concernent la valeur des paramètres d'*AMIGene* (voir la validation automatique p. 250) et la correction manuelle des bornes incorrectes des CDS des banques (voir interface CompAnnotViewer adossée à PkGDB ; voir p. 183).

Parmi les 45 nouveaux gènes potentiels de *S. enterica* serovar Typhimurium LT2 présentés dans

TAB. 10.4 – Nouvelles CDS chez *S. enterica* serovar Typhimurium LT2

Pour les abréviations, voir la légende de la figure 10.9 p. 324.

Bk_label	close to gene	S	begin	end	L	Pc	O	Prot_id	description	Sc	Evalue	id%	comment	M
STM0028.1	*bcfH* /STM0029	D	32116	32445	330	0,74	N	Q9RH97	HYPOTHETICAL 42.7 KDA PROTEIN.	566	2,26E-58	100	STY0034 membrane protein	3
STM0064.1	*dapB* /STM0065	R	74964	75173	210	0,90	Y		NO SIMILARITY				S STM0065	2
STM0291.1	STM0291/STM0292	D	336639	336902	264	0,72	N	CAE15485	HYPOTHETICAL GENE	192	1,81E-14	42		3
STM0294.1	STM0294/STM0295	R	339328	339555	228	0,73	N	P39394	HYPOTHETICAL PROTEIN YJIW.	125	4,45E-07	49	transposase -> FS?	3
STM0342.1	STM0342/STM0343	D	386606	386809	204	0,72	N	Q8Z924	HYPOTHETICAL PROTEIN STY0375.	344	9,58E-33	98	STY0375 Cterm -> FS?	3
STM0398.1	STM0398/STM0399	R	452806	453015	210	0,72	N		NO SIMILARITY					3
STM0412.1	STM0412/*tsx*	D	467346	467453	108	0,72	Y	Q8X4I5	HYPOTHETICAL PROTEIN Z1989.	122	1,15E-06	57	STM0412, STY4175	1
STM0895.1	STM0895/STM0896	R	965099	965353	255	0,84	N	Q83KQ7	HYPOTHETICAL BACTERIOΦ PROTEIN	326	5,13E-30	69	Fels-1 proϕ	2
STM0897.1	STM0897/STM0898	R	967228	967422	195	0,71	N	Q8X359	HYPOTHETICAL PROTEIN ECS1579.	143	3,59E-09	45	Fels-1 proϕ	3
STM0897.2	STM0897/STM0898	R	967415	967768	354	0,84	N		NO SIMILARITY				Fels-1 proϕ	3
STM0910.1	STM0910/STM0911	D	981673	981879	207	0,75	N	Q8X3E1	HYPOTHETICAL PROTEIN ECS0826.	147	1,20E-09	48	Fels-1 proϕ	1
STM0913.1	STM0913/STM0914	D	985514	985861	348	0,86	Y	Q8ZQ93	GIFSY-2 PROΦ, PUTATIVE RECA/RADA RECOMBINASE.	375	5,63E-36	67	STM0913, STM1034 Fels-1 proϕ	1
STM0913.2	STM0913/STM0914	D	985854	986129	276	0,72	Y	Q8ZQ92	GIFSY-2 PROΦ, ATP-BINDING SUGAR TRANSPORTER PROTEIN.	202	7,56E-16	46	STM0913, STM1035 Fels-1 proϕ	1
STM0916.1	STM0916/STM0917	D	987914	988318	405	0,84	N	Q8ZN05	GIFSY-1 PROΦ: SIMILAR TO MINOR TAIL PROTEIN.	486	7,15E-49	72	STM2596 Fels-1 proϕ	1
STM1048.1	STM1048/STM1049	D	1135266	1136141	876	0,85	N	Q8ZN12	GIFSY-1 PROΦ: HOST SPECIFICITY PROTEIN-J IN LAMBDA.	1194	2,10E-131	96	STM2589 Gifsy-2 proϕ	1
STM1123.1	STM1123/*putA*	R	1207597	1208328	732	0,94	N	P10503	Bifunctional putA protein: Proline dehydrogenase (EC 1.5.99.8); Delta-1-pyrroline-5-carboxylate dehydrogenase (EC 1.5.1.12)	1231	7,90E-136	100	STM1124 FS annotated in SP	1
STM1507.1	*ydfJ*/*ydfI*	D	1585527	1585925	399	0,85	N	Q8Z6Z5	Putative membrane transport protein	653	1,95E-68	100	STY1554	1
STM1551.1	STM1551/STM1552	D	1627444	1627776	333	0,73	N	Q8ZFC6	PUTATIVE ACYL CARRIER PROTEIN.	299	5,00E-27	55	STY2213	3
STM1699.1	STM1698A/STM1699	R	1791787	1792194	408	0,68	Y		NO SIMILARITY				S STM1698A STM1699	3
STM1859.1	STM1859/STM1860	R	1956708	1956854	147	0,89	N	O68779	TRANSPOSASE (TRANSPOSASE FOR THE IS285).	228	4,11E-19	85	transposase -> FS	2
STM2008.2	STM2008/*amn*	R	2090460	2090564	105	0,76	N	P77482	PROTEIN YAJQ.	155	1,56E-10	82	FS	2
STM2008.3	STM2008/*amn*	D	2090612	2090785	174	0,70	N		NO SIMILARITY					2
STM2172.1	*yohG* /STM2173	R	2270316	2270474	159	0,72	N		NO SIMILARITY					3
STM2240.1	STM2240/*sspH2*	R	2340564	2340794	231	0,74	N	AAL89446	HYPOTHETICAL 30.2 KDA PROTEIN.	362	7,55E-35	86	ST64B ϕ	2
STM2398.1	*pgtC*/*pgtP*	R	2511738	2511956	219	0,70	N		NO SIMILARITY					3
STM2402.1	*yfdZ*/*glk*	R	2517585	2517746	162	0,87	N		NO SIMILARITY					3
STM2420.1	*xapR*/*xapB*	R	2533330	2533488	159	0,77	N	Q9K6C3	HYPOTHETICAL PROTEIN BH3806.	113	1,24E-05	41		3
STM2507.1	STM2507/STM2508	D	2621728	2621958	231	0,70	N	Q8ZMH8	PUTATIVE CYTOPLASMIC PROTEIN.	170	3,34E-12	44	STM2901	3
STM2614.1	STM2614/STM2615	R	2761369	2761557	189	0,69	N		NO SIMILARITY				STM2615 tRNA, Gifsy-1 proϕ	3
STM2615.1	STM2615/STM2616	R	2761757	2761927	171	0,72	N		NO SIMILARITY				STM2615 tRNA, Gifsy-1 proϕ	3
STM2740.1	STM2740/STM2741	D	2879247	2879414	168	0,76	N	CAC83130	PUTATIVE INTEGRASE INT.	142	4,86E-09	61	Fels-2 proϕ, FS with STM2740	3
STM2798.1	*ygaP*/*stpA*	D	2946953	2947135	183	0,88	N	Q8Z4F3	HYPOTHETICAL PROTEIN STY2919.	210	5,08E-17	95	STY2919	2
STM2954.1	*mazG* /STM2955	R	3102178	3102474	297	0,79	N	Q8XE95	HYPOTHETICAL PROTEIN Z0510.	131	2,04E-07	32		3
STM2614.1	*gudT* /STM2963	R	2761757	2761927	171	0,72	N		NO SIMILARITY					3
STM3083.1	STM3083/STM3084	D	3246649	3246783	135	0,75	N	Q92VL5	mandelate racemase muconate lactonizing enzyme family	109	3,74E-05	58	FS	2
STM3520.1	STM3520/STM3521	R	3682925	3683092	168	0,99	N	Q8ZLH8	PUTATIVE RIBONUCLEOPROTEIN RELATED-PROTEIN.	290	2,15E-26	100	STM3520 pseudotRNA STM3521 Cterm dupplication -> FS	1
STM3521.1	STM3521/*rtcR*	R	3684671	3684793	123	0,84	N		NO SIMILARITY					3
STM3828.1	*dgoA*/*dgoK*	R	4031780	4032076	297	0,49	N	P31458	DgoA: 2-dehydro-3-deoxyphosphogalactonate aldolase (EC 4.1.2.21); Galactonate dehydratase (EC 4.2.1.6)	287	2,80E-25	53,5	FS with dgoA	1
STM3828.2	*dgoA*/*dgoK*	R	4031926	4032396	471	0,85	N	P31458		543	2,09E-55	84	dgoA Nterm -> FS	1
STM4032.1	STM4032/STM4033	D	4241636	4241947	312	0,83	N	Q9KMA6	HYPOTHETICAL PROTEIN VCA0468.	197	4,02E-15	40		3
STM4102.1	STM4102/STM4103	R	4314284	4314490	207	0,78	N	Q9KBJ6	HYPOTHETICAL PROTEIN BH1931.	90	8,10E-02	24,63		3
STM4211.1	STM4211/STM4212	R	4432985	4433113	129	0,82	N	Q8P6H8	Φ-RELATED PROTEIN	68	5,01E+00	43,33 ϕ		2
STM4211.2	STM4211/STM4212	R	4433073	4433243	171	0,75	N	Q9KW60	CAROTOVORUM ORF1, ORF2, ORF4-12 GENES, BACTERIOCIN.	151	4,27E-10	57 ϕ, Cterm STY1628 -> FS		2
STM4211.3	STM4211/STM4212	R	4433269	4433388	120	0,73	N	Q9KW60		113	1,29E-05	54 ϕ, Nterm STY1628 -> FS		2

le tableau 10.4 p. 302, 22 ont été prédits à partir de la matrice des gènes de classe III, 15 tombent dans des régions de bactériophage et 12 sont potentiellement impliqués dans des décalages du cadre de lecture.

10.3.3 *H. influenzae* et *V. cholerae*

Dans le cas de la réannotation du génome d'*H. influenzae*, neuf NG ayant le statut '*newAGC*' codent des fragments protéiques dont sept tombent dans des région où *ProFED* a détecté un décalage du cadre de lecture. Plusieurs de ces fragments sont similaires à des protéines dont la fonction peut être associée à la virulence d'*H. influenzae* (Tab. 10.5 p. 304 ; transport du fer, adhésion, etc.).
Par exemple, en position 1543001 pb du génome, une CDS non annotée présente un décalage du cadre de lecture qui a pour conséquence l'annotation d'une protéine cinq fois plus petite. La protéine correspondante est similaire à la séquence terminale d'une invasine de *Yersinia enterocolitica*.
Un exemple de statut '*ambiguousAGC*' a été trouvé en position 1 481 807 pb du chromosome d'*H. influenzae*, au niveau d'une petite CDS de 249 pb (Pc=0,48) qui chevauche le gène HI1386 annoté comme étant une protéine hypothétique. En utilisant les filtres Seg et Xnu, aucune similitude significative n'a pu être détectée à partir de la protéine correspondante. Sans ces filtres, la séquence a révélé des similitudes dues à la répétition du motif PTNQ au début de la séquence. Nous avons donc regardé plus en détail la séquence nucléique du début de cette CDS : le codon de début de traduction ATG est précédé d'un motif RBS (AAGGAT), et immédiatement suivi de 16 occurences du motif AACC. Il est par conséquent probable que nous ayons à faire à un décalage du cadre de lecture authentique et que ce fragment '*ambiguousAGC*' et HI1386 codent respectivement la partie N-terminale et C-terminale d'un même polypeptide. Il est même possible que l'expression du gène HI1386, localisé à proximité de deux gènes codant des protéines impliquées dans le stockage du fer (*rsgA*) et du gène codant l'antranilate synthase (*trpE*), soit régulée par un mécanisme de *slipped-strand repair*[3] et, par là même, impliquée dans la pathogénie de la bactérie.

Les résultats de réannotation des *deux* chromosomes circulaires (G+C 47%) du pathogène *V. cholerae*, responsable du choléra, sont présentés dans le tableau 10.6 p. 305. L'analyse des résultats montre que la particularité de ce génome est de posséder la proportion la plus importante de GNF ($Pc \geq 0,2$) 2,58% (Tab. 10.1 B p. 290). En effet, ce génome possède une centaine de CDS courtes ($L \leq 150$ pb) et '*probable*' ($0,2 \leq Pc < 0,4$), annotées comme hypothétiques. Elles sont spécifiques de *V. cholerae* et certaines sont similaires entre elles (VC0160 = VC0388 = VC2495 = VC2752 ; G+C 35%). Deux hypothèses sont possibles : (i) ces petites CDS de composition atypique sont orphelines (CDS qui n'ont pas de similitude dans les banques excepté avec des espèces qui sont proches phylogénétiquement) ou (ii) ce sont des artefacts. Il se pourrait que ces duplications aient un lien avec des mécanismes de recombinaison entre les *deux* chromosomes.

[3] Au cours de la réplication, des erreurs sont provoquées dans le nombre de copies des répétitions, ce qui affecte la longueur de la répétition et l'expression du gène (processus aussi appelé *slipped-strand mispairing* ; [Médigue, 2000a]).

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot / CDSa (CDSb)
Begin	S	L (bp)	Name	L (aa)	function	Proba	Id	Name	function	Name	function	
143177	R	135	AFUB_ACTPL	687	FERRIC TRANSPORT SYSTEM PERMEASE	8.7e-15	90	HI0126	ferric ABC transporter, ATP-binding protein (afuC)	HI0129	FERRIC TRANSPORT SYSTEM PERMEASE	AFUB_ACTPL 643 685 / HI0129 1 43
369932	D	183	Q9KLR6	168	IRON-SULFUR CLUSTER-BINDING PROTEIN NAPF	1.2e-13	56	HI0342	ferredoxin-type protein (napF)	HI0343	napD protein (napD)	Q9KLR6 110 162 / HI0342 3 55
615442	R	72	OTCC_HAEIN	334	ORNITHINE Carbamoyl transferase CATABOLIC (EC 2.1.3.3) (OTCASE, ARCB, HI0596)	1.0e-05	100	HI0592	hypothetical protein	HI0594	conserved hypothetical transmembrane protein	OTCC_HAEIN HI0596 311 333 / 1 23
1257573	R	153	Q9K0Q9	219	ALUMINUM RESISTANCE PROTEIN, PUTATIVE.	6.5e-13	66	HI1190	6-pyruvoyl tetrahydro biopterin synthase, putative	HI1191	ALUMINUM RESISTANCE PROTEIN, PUTATIVE.	Q9K0Q9 175 216 / HI1191 5 46
1407129	R	114	Q9ZIX7	216	PUTATIVE TRANSPOSASE	2.2e-12	84	HI1328.1	predicted coding region	HI1329	IS1016-V6 protein (IS1016-V6)	Q9ZIX7 56 93 / HI1329 1 38
1523515	R	147	YQCB_ECOLI	260	Hypothetical 29.7 KDA PROTEIN IN BARA-SYD INTERGENIC REGION	2.6e-09	51	HI1434.2	conserved hypothetical protein	HI1435	conserved hypothetical protein	YQCB_ECOLI 193 239 / HI1435 3 49
1543001	R	237	Q56930	422	YOPA PROTEIN	3.0e-08	39	HI1459	sigma factor, putative	HI1462	conserved hypothetical protein	Q56930 302 379 / 6 78
1686038	R	105	HYPE_ECOLI	322	HYDROGENASE ISOENZYMES FORMATION PROTEIN HYPE	5.9e-06	62	HI1617	aspartate aminotransferase (aspC)	HI1618	ABC transporter, ATP-binding protein	HYPE_ECOLI 288 322 / 1 35
1686795	R	285	Q9S410	205	COBALT MEMBRANE TRANSPORT PROTEIN HOMOLOG (CBIQ)	2.5e-27	68	HI1618	ABC transporter, ATP-binding protein	HI1620	COBALT MEMBRANE TRANSPORT PROTEIN HOMOLOG	Q9S410 118 205 / HI1620 5 92

TAB. 10.5 – CDS *'newAGC'* chez *H. influenzae*

Les trois premières colonnes décrivent la CDS *AMIGene* (début, brin et longueur en paires de bases). Les trois colonnes suivantes décrivent une séquence protéique *subject* résultant d'un Blast2P sur la SWALL (nom de l'entrée, longueur en acides aminés et description fonctionnelle). Les deux colonnes suivantes résument la significativité de l'alignement (E-value et pourcentage d'identité). IF dans la colonne du pourcentage d'identité signifie que le pourcentage est faible à cause d'une région de basse complexité masquée par le filtre Seg. Les deux colonnes suivantes décrivent la CDS annotée qui précède la nouvelle CDS (nom et fonction). Les deux colonnes suivantes décrivent la CDS annotée qui suit la nouvelle CDS (nom et fonction). Enfin la dernière colonne schématise l'alignement.

A

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
323869	D	90	Q9KNI6	46	HYPOTHETICAL PROTEIN VC2753	1.4e-07	90	VC0313	hypothetical protein	VC0314	hypothetical protein	
540598	D	279	TNSB_ECOLI	702	TRANSPOSON TN7 TRANSPOSITION PROTEIN TNSB	5.1e-10	47	VC0511	hypothetical protein	VC0512	methyl-accepting chemotaxis protein	

B

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
272753	R	198	Q9KMR6	722	HYPOTHETICAL PROTEIN VCA0254	2e-15	100	VCA0253	ANTIBIOTIC ACETYLTRANSFERASE	VCA 0254	hypothetical protein	
290611	R	432	NO SIM					VCA0272	hypothetical protein	VCA 0273	hypothetical protein WRONG	
349928	R	240	P95254	83	(RV1960C..) HYPOTHETICAL 9.2 KDA	1e-15	56	VCA0359	PLASMID STABILIZATION ELEMENT PARE, PUTATIVE.	VCA 0360	hypothetical protein	
362854	R	132	Q9KMC8	44	HYPOTHETICAL PROTEIN VCA0434	6e-18	95	VCA0380	HYPOTHETICAL PROTEIN	VCA 0381	hypothetical protein	
432125	D	123	Q9KMM9	60	HYPOTHETICAL PROTEIN VCA0302	2e-10	73	VCA0500	HYPOTHETICAL PROTEIN WRONG	VCA 0501	hypothetical protein	
699968	R	435	NO SIM					VCA0755	HYPOTHETICAL PROTEIN WRONG WRONG	VCA 0756	TRANSCRIPTIONAL REGULATOR, LYSR FAMILY	

TAB. 10.6 – CDS *'newAGC'* chez *V. cholerae*

A) Chromosome I.

B) Chromosome II. Pour la légende, se référer à celle du tableau 10.5 p. 304.

10.3.4 *N. meningitidis*

Diversité du processus d'annotation dans le cas des doublons

Le séquençage de « doublons » (deux espèces du même genre ou deux souches d'une même espèce) devrait nous permettre d'affiner les annotations de ces génomes grâce à des comparaisons (génomes des *Chlamydiae*, des deux souches d'*H. pylori*, des *Mycoplasmes*, des *Pyrococccus*). Pour ces quatre doublons, il y a au minimum un an de différence entre les deux publications et, à chaque fois, l'annotation la plus récente révèle moins de GNF '*wrongBank*' et moins de NG '*newAGC*' :

- *C. trachomatis* (1998) vs *Chlamydia pneumoniae* (1999) : 0,34 vs 0,09% GNF '*wrongBank*' et 0,77 vs 0,56% NG '*newAGC*' ;
- *H. pylori* 26695 (1997) vs *H. pylori* J99 (1999) : 1,89 vs 0,13% GNF '*wrongBank*' et 0,70 vs 0,40% NG '*newAGC*' ;
- *M. genitalium* (1995) vs *M. pneumoniae* (1996) : 0,41 vs 0,29% GNF '*wrongBank*' et 6 vs 4,60% NG '*newAGC*' ;
- *Pyrococcus horikoshii* (1997) vs *P. abyssi* (1999) : 14,67 vs 0,00% GNF '*wrongBank*' et 2,76 vs 1,45% NG '*newAGC*'.

Les deux souches de *N. meningitidis* méritent une attention particulière. Ce sont des génomes difficiles à annoter parce qu'ils ont un contenu en G+C très variable (déviation standard du G+C des CDS de longueur L>300 bp égale à 6,9% ; Tab. 10.3 p. 297 [Hsiao *et al.*, 2003]), et aussi parce que leur grande plasticité génère de nombreuses répétitions, remaniements et décalages du cadre de lecture. Les difficultés d'annotation de CDS chez ces génomes en particulier se résument par :

- la prédiction de CDS dans les régions de composition atypique (généralement A+T riches ; *sacB* Fig. 7.1 B p. 198),
- la prédiction de CDS dans les régions contenant des CDS fantômes (la région répétée de NMA1353 génère la CDS fantôme NMA1353.1 ; Fig. 7.1 A p. 198),
- la prédiction de CDS dans les régions altérées contenant des CDS fragmentées et recouvrantes (*e.g.* décalages du cadre de lecture générant le pseudogène composé des fragments NMA1359.1 et NMA1360.1 ; Fig. 7.1 A p. 198) ou
- le réajustement du codon d'initiation, d'autant plus important que le génome est G+C riche (NMA1359.1 ; Fig. 7.1 A p. 198).

Les annotations de ces deux génomes ont été publiées en 2000. Nous avons assisté à une course au séquençage plutôt qu'à une collaboration fructueuse entre les deux plus grands centres internationaux de séquençage (le Sanger Center au Royaume-Uni et le TIGR aux Etats-unis). La première version du fichier d'annotation de *N. meningitidis* MC58 (serogroup B) du TIGR datait du 13 mars 2000, puis la version actuelle du fichier d'annotation de *N. meningitidis* Z2491 (Serogroup A) du Sanger est arrivée, datée du 30 mars 2000. Ensuite, sans publication, une seconde version de *N. meningitidis* MC58 (serogroup B) a remplacé la première ; cette version actuelle date du 5 avril 2000. Par ce stratagème, le TIGR a réussi à publier dans Science le 10 mars 2000, avant le Sanger qui a publié dans Nature le 30 mars 2000. Nous avons réannoté les deux versions de *N. meningitidis*

MC58 (serogroup B). La différence la plus flagrante est que le nombre d'OA est passé de 2035 à 2128, ce qui a eu pour conséquence de faire chuter le nombre de NG '*newAGC*' de 128 à 50 (Tab. 10.4 p. 308). Malgré la mise à jour du TIGR, les résultats de *N. meningitidis* MC58 (serogroup B) sont moins bons que ceux de *N. meningitidis* Z2491 (Serogroup A), par exemple dans le nombre de GNF de faible Pc et surtout dans le nombre de NG de forte Pc (Tab. 10.1 B p. 290).
Sachant que les prédictions obtenues par les deux plus grands centres de séquençage sont généralement justes (*e.g. C. jejuni* pour le Sanger et *H. influenzae* pour le TIGR ; Tab. 10.1 p. 290), la proportion de GNF de faible Pc est surprenante : 240 soit 11,63% chez *N. meningitidis* Z2491 (Serogroup A) et 308 soit 14,47% chez *N. meningitidis* MC58 (serogroup B). En terme de processus d'annotation, les stratégies utilisées par ces deux groupes sont globalement semblables, même s'ils utilisent leurs propres outils[4]. Dans les cas de *Aeropyrum pernix* et *P. horikoshii*, qui ont les plus forts pourcentages de GNF ($Pc < 0,2$), nous avons vu que la majorité de ces GNF se voyaient attribuer le statut '*wrongBank*'. Au contraire, pour les souches *N. meningitidis* Z2491 (Serogroup A) et *N. meningitidis* MC58 (serogroup B) qui ont les plus forts pourcentages de GNF de faible Pc après *A. pernix* et *P. horikoshii*, ce phénomène n'est pas observé. En effet la majorité des GNF ($Pc < 0,2$) se retrouvent alors sans statut ('*noStatusBank*'). Nous concluons cette fois-ci que la grande proportion de GNF de faible Pc n'est pas due aux processus d'annotation mais à des propriétés intrinsèques des séquences génomiques.

Impact des gènes de composition A+T riche et des répétitions

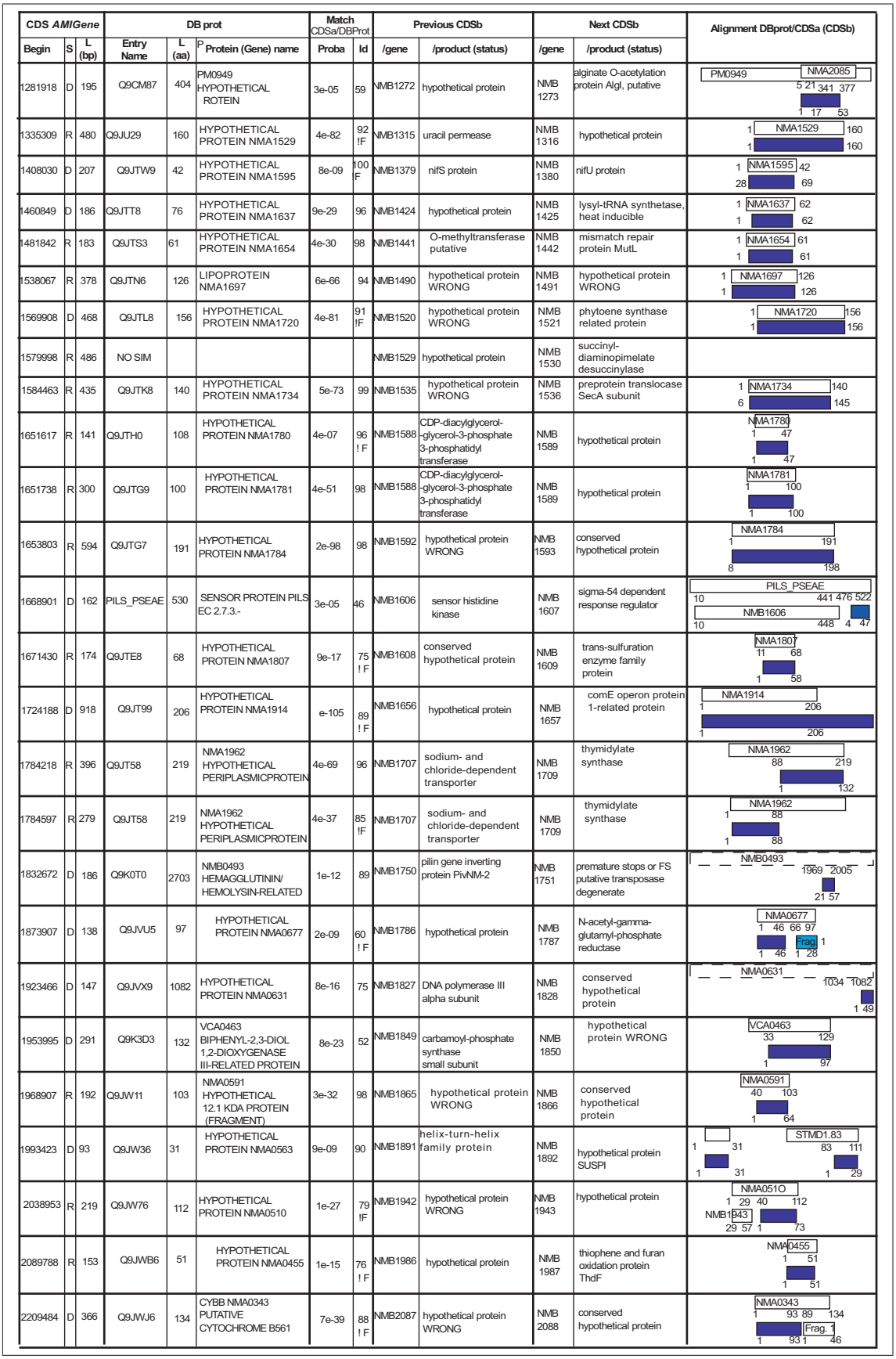
Deux hypothèses peuvent expliquer les propriétés particulières des chromosomes des deux souches de *N. meningitidis* générant une grande proportion de GNF de faible Pc :

1. Il existe différentes classes de gènes dans l'usage des codons synonymes. Certaines régions contiennent des gènes qui utilisent préférentiellement les codons se terminant par A ou T. Le non-codant étant plus riche en AT que le codant, ces régions codantes sont mal détectées par les programmes de prédiction qui les confondent avec du non-codant. Une prédiction efficace de ce type de gènes nécessite un apprentissage spécifique de cette classe de gènes (une matrice de transition supplémentaire spécifique de cette classe). L'explication biologique triviale associée à ce biais dans l'usage des codons synonymes est celle du transfert horizontal. Par exemple, les souches de *Neisseria* sont naturellement compétentes. Elles peuvent capturer de l'ADN dans l'environnement et l'incorporer dans leur génome. Il est donc possible que de larges portions du génome aient été acquises par un mécanisme de transfert horizontal. On ne peut nier l'existence d'une classe de gènes A+T riches mais l'origine exogène de ces gènes n'est pas démontrée, juste suggérée. Le débat sur le transfert horizontal, les biais dans l'usage des codons synonymes et l'évolution des génomes procaryotes est d'actualité ; c'est un sujet complexe et loin d'être résolu. L'exploration des génomes procaryotes et les expérimentations

[4]La stratégie du TIGR paraît plus complexe et plus automatique que celle du Sanger. Une plus grande expertise manuelle est perçue dans les résultats du Sanger, qui sont de ce fait plus difficiles à analyser.

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
113210	R	168	Q9JWY4	52	NMA0170 PUTATIVE LIPOPROTEIN	3e-23	96	NMB0104	HYPOTHETICAL PROTEIN	NMB 0105	PHNO-RELATED PROTEIN	NMA0170 1 52 / 5 56
264695	D	285	Q9JSL9	116	NMA2227 HYPOTHETICAL INNER MEMBRANE PROTEIN	4e-49	92	NMB0260	hypothetical protein WRONG	NMB 0261	geranyltranstransferase	NMA2227 22 116 / 1 95
274394	D	141	Q9JSN2	208	HYPOTHETICAL PROTEIN NMA2214	2e-09	87	NMB0271	hypothetical protein	NMB 0272	hypothetical protein WRONG	NMA2214 35 66 / Frag. 1 / NMB0273 7 38
315669	R	90	Q9JSQ6	30	NMA2182 HYPOTHETICAL PROTEIN	8e-10	96	NMB0304	opacity pseudogene	NMB 0305	hypothetical protein WRONG	NMA2182 1 30 / 1 30
607688	D	183	Q9JVN7	39	HYPOTHETICAL PROTEIN NMA0764	2e-14	100	NMB0579	copper ABC transporter, ATP-binding protein	NMB 0580	protein disulfide isomerase NosL putative	NMA0764 1 39 / 23 61
659026	R	75	Q9JVH8	24	NMA0835 PEPTIDE CHAIN RELEASE FACTOR 3 REMNANT (FRAGMENT)	3e-06	95	NMB0625	conserved hypothetical protein	NMB 0626	peptide chain release factor 3	NMA0835 4 24 / 1 21
704462	D	195	Q9JVE1	54	HYPOTHETICAL PROTEIN NMA0877	3e-24	98	NMB0676	HYPOTHETICAL PROTEIN	NMB 0677	HYPOTHETICAL PROTEIN WRONG	NMA0877 1 54 / 12 65
746077	R	384	Q9KL19	341	HYPOTHETICAL PROTEIN VCA0930	6e-18	43	NMB0714	conserved hypothetical protein	NMB 0715	HYPOTHETICAL PROTEIN	VCA0930 31 142 207 331 / NMB0714 1 112 1 116
761609	D	345	Q9JV99	96	HYPOTHETICAL PROTEIN NMA0939	2e-42	85 !F	NMB0729	integration host factor, alpha subunit	NMB 0730	HYPOTHETICAL PROTEIN	NMA0939 1 96 / 20 115
770873	D	225	Q9JV88	200	HYPOTHETICAL PROTEIN NMA0951	2e-12	100 !F	NMB0739	conserved hypothetical protein	NMB 0740	DNA repair protein RecN	NMA0950 / NMA0951 1 309 1 133 126 200 / NMB0739 1 309 348 480 1 75
790027	D	459	Q9JV68	201	NMA0973 PUTATIVE INTEGRAL MEMBRANE PROTEIN	7e-82	94 !F	NMB0761	HYPOTHETICAL PROTEIN WRONG	NMB 0762	HYPOTHETICAL PROTEIN WRONG	NMA0973 49 201 / 1 153
873544	R	381	Q9JV05	230	HYPOTHETICAL PROTEIN NMA1058	1e-56	85 !F	NMB0846	LPS biosynthesis protein-related protein	NMB 0847	hypothetical protein	NMA1058 60 175 / 1 116
930992	R	165	Q9JVR0	55	NMA0735 HYPOTHETICAL PROTEIN	3e-25	98	NMB0919	IS1106 transposase putative	NMB 0920	isocitrate dehydrogenase	NMA0735 1 55 / 1 55
1020182	R	177	Q9JUP5	60	NMA1208 HYPOTHETICAL PROTEIN	3e-26	93	NMB0999	NifR3/SMM1 family protein	NMB 1000	This premature stops or frameshifts	NMA1208 1 60 / 1 59
1021817	R	237	Q9JUN7	67	NMA1218 HYPOTHETICAL PROTEIN	4e-32	92	NMB1000	This premature stops or frameshifts	NMB 1001	integrase protein, degenerate premature stops or frameshifts	NMA1218 1 67 / 10 79
1022041	R	279	Q9JUN6	108	NMA1219 HYPOTHETICAL PROTEIN	3e-50	96	NMB1000	This premature stops or frameshifts	NMB 1001	integrase protein, degenerate premature stops or frameshifts	NMA1219 16 108 / 1 93
1022360	R	180	Q9JUN5	87	NMA1220 HYPOTHETICAL PROTEIN	3e-05	41 !F	NMB1000	This premature stops or frameshifts	NMB 1001	integrase protein, degenerate premature stops or frameshifts	NMA1220 28 87 / 1 60
1106096	D	162	Q9JV05	230	NMA1189 DUPLICATED HYPOTHETICAL PROTEIN	7e-15	72 !F	NMB1085	N-acetylmuramoyl-L-alanine amidase putative	NMB 1086	hypothetical protein	NMA1189 1 54 / 1 54
1106471	D	417	Q9JUQ5	139	HYPOTHETICAL PROTEIN NMA1191	1e-60	84 !F	NMB1086	hypothetical protein	NMB 1087	hypothetical protein	NMA1191 1 139 / 1 139
1145282	R	555	Q9JUF1	195	HYPOTHETICAL PROTEIN NMA1345	7e-96	95	NMB1135	hypothetical protein	NMB 1136	hypothetical protein	NMA1345 2 181 / 1 180
1154339	D	1047	NO SIM					NMB1147	hypothetical protein	NMB 1186	hypothetical protein	
1177318	R	555	Q9JUF1	195	HYPOTHETICAL PROTEIN NMA1345	9.8e-89	95	NMB1173	hypothetical protein	NMB 1174	hypothetical protein	NMA1345 1 184 / 2 185
1186375	D	1047	NO SIM					NMB1185	hypothetical protein	NMB 1186	hypothetical protein	
1252211	D	432	Q9JUA8	144	Putative periplasmic protein NMA1414	2e-65	88 !F	NMB1245	hypothetical protein	NMB 1246	conserved hypothetical protein	1 NMA1414 144 / 1 144

FIG. 10.4 – CDS '*newAGC*' chez *N. meningitidis* MC58 (serogroup B) (pour la légende voir TAB. 10.5 p. 304)



CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	P Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
1281918	D	195	Q9CM87	404	PM0949 HYPOTHETICAL ROTEIN	3e-05	59	NMB1272	hypothetical protein	NMB 1273	alginate O-acetylation protein Algl, putative	PM0949 NMA2085 5 21 341 377 1 17 53
1335309	R	480	Q9JU29	160	HYPOTHETICAL PROTEIN NMA1529	4e-82	92 !F	NMB1315	uracil permease	NMB 1316	hypothetical protein	1 NMA1529 160 1 160
1408030	D	207	Q9JTW9	42	HYPOTHETICAL PROTEIN NMA1595	8e-09	100 F	NMB1379	nifS protein	NMB 1380	nifU protein	1 NMA1595 42 28 69
1460849	D	186	Q9JTT8	76	HYPOTHETICAL PROTEIN NMA1637	9e-29	96	NMB1424	hypothetical protein	NMB 1425	lysyl-tRNA synthetase, heat inducible	1 NMA1637 62 1 62
1481842	R	183	Q9JTS3	61	HYPOTHETICAL PROTEIN NMA1654	4e-30	98	NMB1441	O-methyltransferase putative	NMB 1442	mismatch repair protein MutL	1 NMA1654 61 1 61
1538067	R	378	Q9JTN6	126	LIPOPROTEIN NMA1697	6e-66	94	NMB1490	hypothetical protein WRONG	NMB 1491	hypothetical protein WRONG	1 NMA1697 126 1 126
1569908	D	468	Q9JTL8	156	HYPOTHETICAL PROTEIN NMA1720	4e-81	91 !F	NMB1520	hypothetical protein WRONG	NMB 1521	phytoene synthase related protein	1 NMA1720 156 1 156
1579998	R	486	NO SIM					NMB1529	hypothetical protein	NMB 1530	succinyl-diaminopimelate desuccinylase	
1584463	R	435	Q9JTK8	140	HYPOTHETICAL PROTEIN NMA1734	5e-73	99	NMB1535	hypothetical protein WRONG	NMB 1536	preprotein translocase SecA subunit	1 NMA1734 140 6 145
1651617	R	141	Q9JTH0	108	HYPOTHETICAL PROTEIN NMA1780	4e-07	96 ! F	NMB1588	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyl transferase	NMB 1589	hypothetical protein	NMA1780 1 47 1 47
1651738	R	300	Q9JTG9	100	HYPOTHETICAL PROTEIN NMA1781	4e-51	98	NMB1588	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyl transferase	NMB 1589	hypothetical protein	NMA1781 1 100 1 100
1653803	R	594	Q9JTG7	191	HYPOTHETICAL PROTEIN NMA1784	2e-98	98	NMB1592	hypothetical protein WRONG	NMB 1593	conserved hypothetical protein	NMA1784 1 191 8 198
1668901	D	162	PILS_PSEAE	530	SENSOR PROTEIN PILS EC 2.7.3.-	3e-05	46	NMB1606	sensor histidine kinase	NMB 1607	sigma-54 dependent response regulator	PILS_PSEAE 10 441 476 522 NMB1606 10 448 4 47
1671430	R	174	Q9JTE8	68	HYPOTHETICAL PROTEIN NMA1807	9e-17	75 ! F	NMB1608	conserved hypothetical protein	NMB 1609	trans-sulfuration enzyme family protein	NMA1807 11 68 1 58
1724188	D	918	Q9JT99	206	HYPOTHETICAL PROTEIN NMA1914	e-105	89 ! F	NMB1656	hypothetical protein	NMB 1657	comE operon protein 1-related protein	NMA1914 1 206 1 206
1784218	R	396	Q9JT58	219	NMA1962 HYPOTHETICAL PERIPLASMICPROTEIN	4e-69	96	NMB1707	sodium- and chloride-dependent transporter	NMB 1709	thymidylate synthase	NMA1962 88 219 1 132
1784597	R	279	Q9JT58	219	NMA1962 HYPOTHETICAL PERIPLASMICPROTEIN	4e-37	85 !F	NMB1707	sodium- and chloride-dependent transporter	NMB 1709	thymidylate synthase	NMA1962 1 88 1 88
1832672	D	186	Q9K0T0	2703	NMB0493 HEMAGGLUTININ/ HEMOLYSIN-RELATED	1e-12	89	NMB1750	pilin gene inverting protein PivNM-2	NMB 1751	premature stops or FS putative transposase degenerate	NMB0493 1969 2005 21 57
1873907	D	138	Q9JVU5	97	HYPOTHETICAL PROTEIN NMA0677	2e-09	60 ! F	NMB1786	hypothetical protein	NMB 1787	N-acetyl-gamma-glutamyl-phosphate reductase	NMA0677 1 46 66 97 Frag 1 1 46 1 28
1923466	D	147	Q9JVX9	1082	HYPOTHETICAL PROTEIN NMA0631	8e-16	75	NMB1827	DNA polymerase III alpha subunit	NMB 1828	conserved hypothetical protein	NMA0631 1034 1082 1 49
1953995	D	291	Q9K3D3	132	VCA0463 BIPHENYL-2,3-DIOL 1,2-DIOXYGENASE III-RELATED PROTEIN	8e-23	52	NMB1849	carbamoyl-phosphate synthase small subunit	NMB 1850	hypothetical protein WRONG	VCA0463 33 129 1 97
1968907	R	192	Q9JW11	103	NMA0591 HYPOTHETICAL 12.1 KDA PROTEIN (FRAGMENT)	3e-32	98	NMB1865	hypothetical protein WRONG	NMB 1866	conserved hypothetical protein	NMA0591 40 103 1 64
1993423	D	93	Q9JW36	31	HYPOTHETICAL PROTEIN NMA0563	9e-09	90	NMB1891	helix-turn-helix family protein	NMB 1892	hypothetical protein SUSPI	STMD1.83 1 31 83 111 1 31 1 29
2038953	R	219	Q9JW76	112	HYPOTHETICAL PROTEIN NMA0510	1e-27	79 !F	NMB1942	hypothetical protein WRONG	NMB 1943	hypothetical protein	NMA0510 1 29 40 112 NMB1943 29 57 1 73
2089788	R	153	Q9JWB6	51	HYPOTHETICAL PROTEIN NMA0455	1e-15	76 ! F	NMB1986	hypothetical protein	NMB 1987	thiophene and furan oxidation protein ThdF	NMA0455 1 51 1 51
2209484	D	366	Q9JWJ6	134	CYBB NMA0343 PUTATIVE CYTOCHROME B561	7e-39	88 ! F	NMB2087	hypothetical protein WRONG	NMB 2088	conserved hypothetical protein	NMA0343 1 93 89 134 Frag. 1 1 93 1 46

FIG. 10.4 – CDS '*newAGC*' chez *N. meningitidis* MC58 (serogroup B)

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
257463	R	198	Q9K1Q2 Q9JUW2	72 160	NMB0028 hypothetical prot NMA1111 hypothetical prot	3e-16 3e-04	93 95	NMA0273	FKBP_NEIMA FK506-BINDING PROTEIN	NMA 0274	PUTATIVE GLYCERATE DEHYDROGENASE (EC 1.1.1.29)	
969210	R	258	Q9K030	120	HYPOTHETICAL PROTEIN NMB0794	5e-40	89	NMA1003	PUTATIVE TRANSMEMBRANE TRANSPORT PROTEIN	NMA 1004	PEPTIDYL-TRNA HYDROLASE (EC 3.1.1.29) (PTH)	
1064496	R	198	Q9JVR0	55	NMA0735 HYPOTHETICAL PROTEIN	3e-25	98	NMA1115	TRANSPOSASE FOR IS1106A3	NMA 1116	ICD ISOCITRATE DEHYDROGENASE (EC 1.1.1.42)	
1181314	R	270	Q9JZM0	370	NMB0991 IS1106 TRANSPOSASE	8e-47	97	NMA1253	TRANSPOSASE FOR IS1106A3	NMA 1254	SERINE HYDROXYMETHYL TRANSFERASE	
1207425	R	168	Q9JTC1	234	NMA1884 PUTATIVE REGULATOR	1e-15	62	NMA1281	HYPOTHETICAL PROTEIN artefact	NMA 1282	PUTATIVE DNA-BINDING PROTEIN	
1256374	D	1047	NO SIM					NMA1359	HYPOTHETICAL PROTEIN	NMA 1360	HYPOTHETICAL PROTEIN WRONG	
1370874	D	195	Q9CM87	404	PM0949 HYPOTHETICAL PROTEIN	2e-05	56	NMA1476	putative mercuric ion binding protein	NMA 1478	putative polysaccharide modification protein	
1598048	R	339	Q9JZM0	370	NMB0991 IS1106 TRANSPOSASE	2e-59	98	NMA1674	PUTATIVE TRANSGLYCOSYLASE	NMA 1675	HYPOTHETICAL PROTEIN	
1603546	R	282	Q9JYR2	107	NMB1468 HYPOTHETICAL PROTEIN	7e-18	52 ! F	NMA1680	PUTATIVE MEMBRANE PROTEIN	NMA 1681	HYPOTHETICAL PROTEIN	
1661936	R	516	NO SIM					NMA1729	HYPOTHETICAL OUTER MEMBRANE PROTEIN	NMA 1730	SUCCINYL-DIAMINOPIMELATE DESUCCINYLASE (EC 3.5.1.18). dapE	
1750109	D	270	PILS_PSEAE	530	SENSOR PROTEIN PILS EC 2.7.3.-	3e-05	46	NMA1803	PUTATIVE TWO COMPONENT SENSOR KINASE	NMA 1805	PUTATIVE TWO COMPONENT RESPONSE REGULATOR	
1936386	D	234	Q9JY52	117	NMB1739 HYPOTHETICAL PROTEIN	3e-38	94	NMA1996	periplasmic type I secretion system protein (natD')	NMA 1997	Hypothetical protein	

TAB. 10.7 – CDS '*newAGC*' chez *N. meningitidis* Z2491 (Serogroup A) (pour la légende voir TAB. 10.5 p. 304)

biologiques devraient aider à la compréhension du mécanisme de transfert horizontal, qui semble jouer dans l'histoire des génomes un rôle bien plus important qu'on l'avait jusqu'alor imaginé (voir p. 333).

2. La plasticité génomique est la caractéristique la plus frappante de *N. meningitidis*. En effet, on trouve de nombreuses répétitions de toutes sortes, allant des petites répétitions (2000 copies de la séquence *DNA uptake* impliquée dans la reconnaissance et la capture d'ADN dans l'environnement) aux séquences d'insertion présentant des mutations (décalages du cadre de lecture, délétions, codons stop prématurés) en passant par les duplications de gènes d'un kilobase ou plus. L'explication biologique associée cette hypothèse est celle de l'évolution rapide permettant une grande variabilité antigénique de ce pathogène des muqueuses [Parkhill *et al.*, 2000, Saunders *et al.*, 2000], comme nous l'avons déjà vu pour les mycoplasmes.

Finalement, les différents niveaux de plasticité peuvent engendrer des conflits d'annotation entre deux génomes d'une même espèce (répétitions à l'intérieur d'une CDS à l'origine d'une CDS fantôme, CDS dupliquées dont une copie peut être fragmentée), voire même à l'intérieur d'un génome (duplication de régions chromosomiques[5] ; Tab. 10.7 p. 310 et Tab. 10.4 p. 308).

Améliorations et conclusions

La progression des résultats la plus spectaculaire est celle associée au génome de *N. meningitidis* Z2491 (Serogroup A) ; on passe en effet de 87,35 % de CDS communes (ligne NEIMA* ; Tab. 10.2 p. 295) à 96,37% (ligne NEIMA). Ceci est probablement dû à l'utilisation de matrices spécifiques de l'usage des codons synonymes qui permettent de mieux prédire les régions atypiques, comme le montre la figure 7.1 B p. 198.

Dans le cas de *N. meningitidis* MC58 (serogroup B), la progression est moins spectaculaire (85,06 à 92,85% de CDS communes). Les résultats de réannotation entre *N. meningitidis* Z2491 (Serogroup A) et *N. meningitidis* MC58 (serogroup B) sont différents (*e.g.* resp. 0,28–1,85 % CDS '*wrongBank*' et 0,70–2,39 % CDS '*newAGC*' ; Tab. 10.2 p. 295) alors que les deux chromosomes présentent plus de 90% d'identité au niveau nucléique [Guibourdenche *et al.*, 1986]. L'hypothèse la plus probable pour expliquer ce résultat inattendu est une hétérogénéité d'annotation. En effet, dans le cas des résultats de la ligne NEIMA* du tableau 10.2 p. 295, le pourcentage faible de CDS communes s'expliquait en grande partie par la présence de régions de composition atypique dont la prédiction nécessite une matrice spécifique. Alors que, dans le cas de la ligne NEIMB*, il semble persister un facteur supplémentaire : un nombre non négligeable de CDS mal prédites (Fig. 10.5 p. 312).

Dans des séquences d'ADN portant des répétitions, la proportion de faux-positifs détectée par certaines méthodes fondées sur les chaînes de Markov peut être élevée. Effectivement, les CDS de *N. meningitidis* MC58 (serogroup B) ayant le statut '*wrongBank*' sont annotées comme protéine

[5]Par exemple, chez *N. meningitidis* MC58 (serogroup B), les CDS '*newAGC*' NMB1085.1, NMB0989.1 et NMB1086.1 sont respectivement similaires aux CDS NMB0989, NMB1086 et NMB0990.

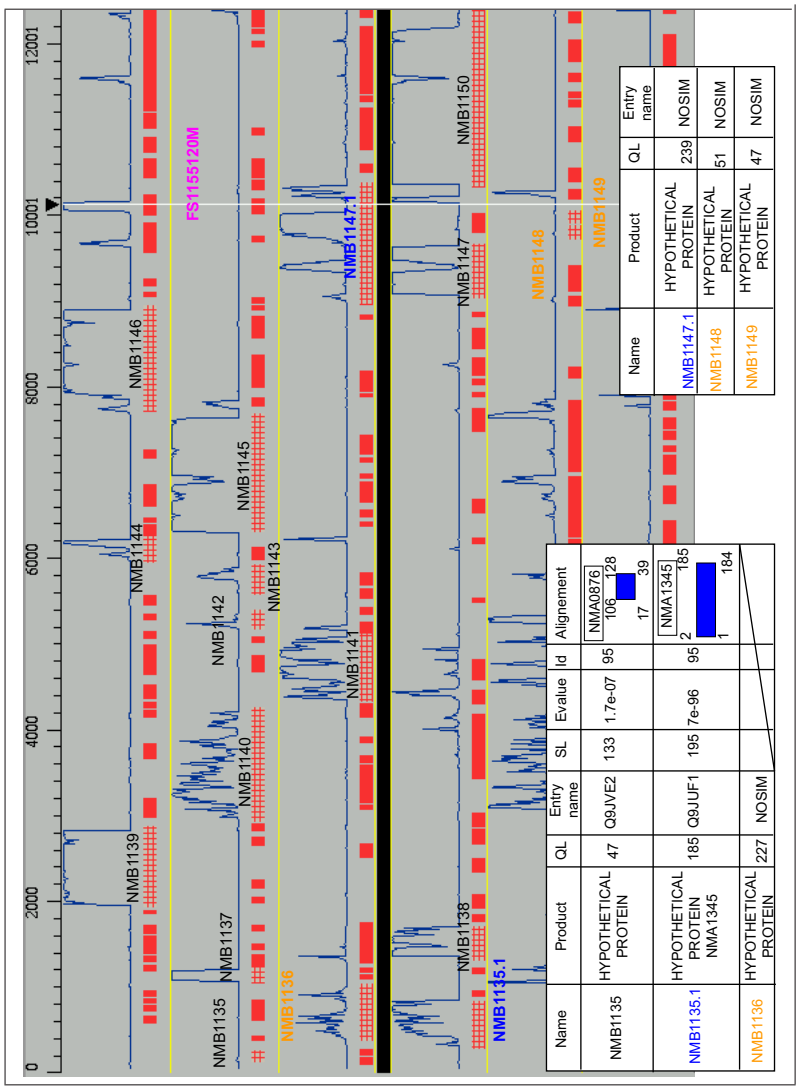
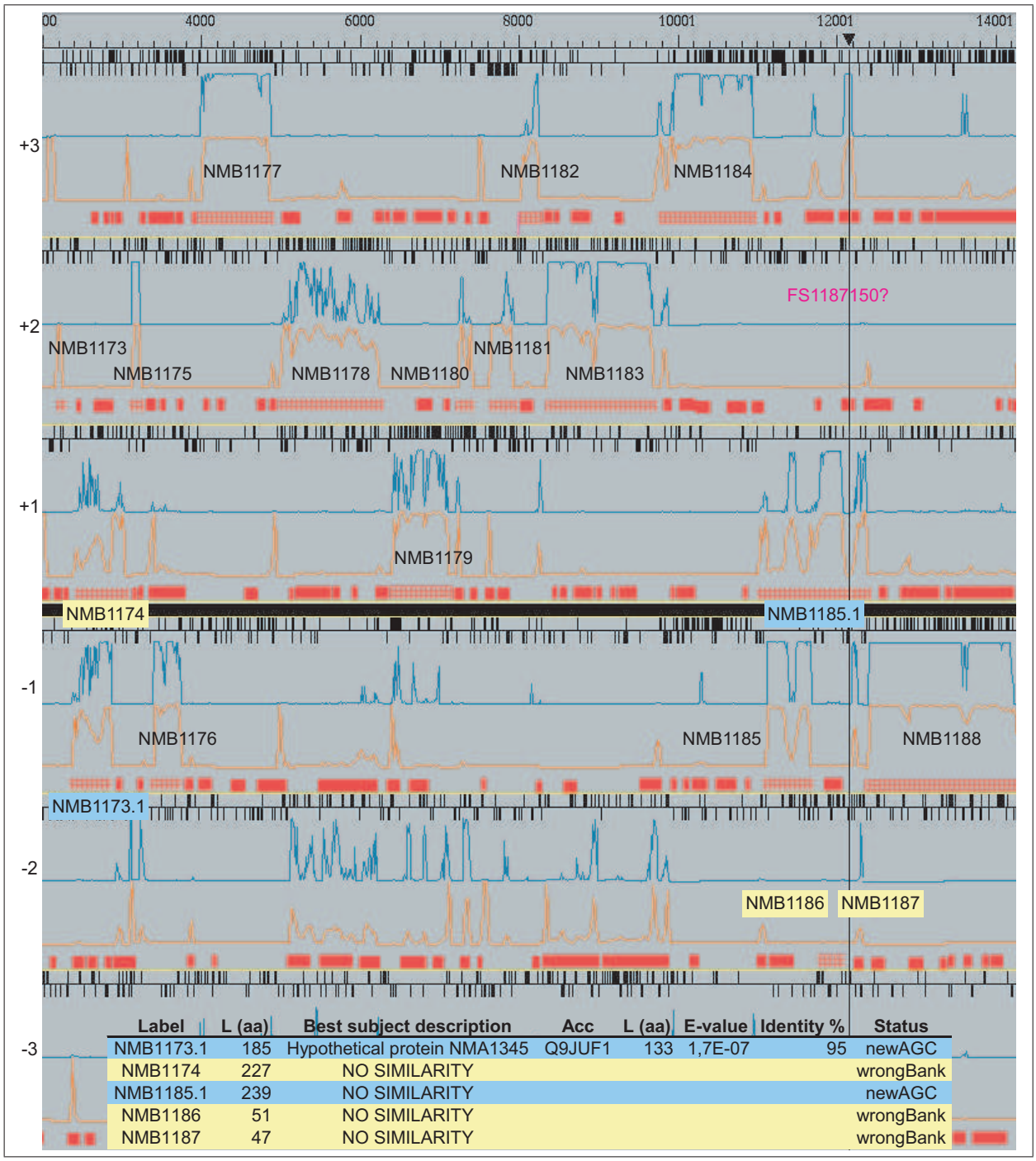


FIG. 10.5 – A) CDS '*newAGC*' et '*wrongBank*' chez *N. meningitidis* MC58 (serogroup B)

Visualisation des six phases de la région 1145-1163 du chromosome de *N. meningitidis* MC58 (serogroup B) dans le logiciel Imagene. Cette figure représente une CDS '*wrongBank*' NMB1136 face à une CDS '*newAGC*' NMB1135.1 et deux CDS '*wrongBank*' NMB1148 et NMB1149 face à une CDS '*newAGC*' NMB1147.1. La séquence protéique de 185 aa de NMB1135.1 présente une forte similitude avec NMA1345, protéine hypothétique de 195 aa. NMB1135.1 est donc une protéine complète, hypothétique et conservée.

La séquence protéique de 239 aa de NMB1147.1 (si on réajuste le codon d'initiation de manière à éliminer le chevauchement avec NMB1147) ne présente pas de similitude significative. Cependant elle présente une faible similitude (E-value=0,024) avec la protéine capA (biosynthèse de la capsule Q8F6F6) de *Leptospira interrogans* et avec la protéine Tll1801 (Q8DHZ2) de *Synechococcus elongatus*. Nous retrouvons son orthologue NMA1359.1 ayant aussi le statut '*newAGC*'. NMB1147.1 est donc une protéine hypothétique. Elle présente un cas de décalage du cadre de lecture compensé. Le petit fragment codant inclus dans NMB1147.1 présente une similarité significative (E-value=1,2 e-06) avec les mêmes protéines capA et Tll1801. Ce décalage de phase est retrouvé chez *N. meningitidis* Z2491 (Serogroup A). Cela ne semble donc pas être une erreur de séquençage. NMB1147.1 semble être un vestige de protéine (pseudogène dont la séquence est dégénérée) composé de deux fragments. Nous sommes donc dans une région bousculée difficile à annoter qui peut être impliquée dans la virulence des *Neisseria*. Plusieurs indicateurs (probabilité et similitude) permette d'écarter l'hypothèse que NMB1147.1 serait une CDS artefactuelle (un faux-positif).

Cette région est dupliquée chez *N. meningitidis* MC58 (serogroup B) mais pas chez *N. meningitidis* Z2491 (Serogroup A). Chez *N. meningitidis* Z2491 (Serogroup A), on ne trouve qu'une seule région de NMA1332 à NMA1359.1. Chez *N. meningitidis* Z2491 (Serogroup A), la protéine NMA1357 a un statut '*suspiciousBank*'. Elle équivaut à la CDS entre NMB1145 et NMB1146 sur la même phase que NMB1145. Chez *N. meningitidis* Z2491 (Serogroup A), la protéine NMA1360 a un statut '*wrongBank*'. Elle équivaut à la CDS entre NMB1147 et NMB1150 sur la même phase. Ainsi on trouve un désaccord entre l'annotation du TIGR, du Sanger et *AMIGene*. Avec cet exemple, on voit bien que le fait que deux CDS soient conservées entre 2 espèces proches n'est pas suffisant pour dire qu'elles sont avérées. On doit se méfier des similitudes de séquences entre espèces proches, surtout quand l'annotation d'une des deux espèces a été calquée sur l'autre. L'effet boule de neige des erreurs d'annotation est pervers ; en effet, la propagation d'erreurs d'annotation empêche de les mettre en évidence sur la base des résultats de similitudes, car ces protéines présentent alors des similitudes du type *protéine hypothétique conservée*.



Label	L (aa)	Best subject description	Acc	L (aa)	E-value	Identity %	Status
NMB1173.1	185	Hypothetical protein NMA1345	Q9JUF1	133	1,7E-07	95	newAGC
NMB1174	227	NO SIMILARITY					wrongBank
NMB1185.1	239	NO SIMILARITY					newAGC
NMB1186	51	NO SIMILARITY					wrongBank
NMB1187	47	NO SIMILARITY					wrongBank

FIG. 10.5 – B) CDS '*newAGC*' et '*wrongBank*' chez *N. meningitidis* MC58 (serogroup B)

Visualisation des six phases de la région 1175-1193 du chromosome de *N. meningitidis* MC58 (serogroup B) dans le logiciel Imagene. Cette figure représente une CDS '*wrongBank*' NMB1174 face à une CDS '*newAGC*' NMB1173.1 et deux CDS '*wrongBank*' NMB1186 et NMB1187 face à une CDS '*newAGC*' NMB1185.1. Les deux régions de NMB1123 à NMB1159 et de NMB1161 à NMB1197 sont dupliquées. Par exemple, on a respectivement NMB1135, NMB1135.1, NMB1136, NMB1147.1, NMB1148 et NMB1149 qui sont identiques à NMB1173, NMB1173.1, NMB1174, NMB1185.1, NMB1186, et NMB1187.

Légende : les rectangles rouges sont les CDS de longueur maximale, les rectangles rouge quadrillés sont les CDS annotées ou prédites par *AMIGene*, les courbes bleues et oranges sont respectivement les probabilités de codage prédites par *GeneMark* et *GLIMMER* 1.01. Les noms de CDS en noir représente des CDS communes aux annotations et à la stratégie *AMIGene*. Les noms de CDS en jaune représentent des CDS uniques aux annotations des banques ayant le statut '*wrongBank*'. Le nom de CDS en bleu représente une CDS unique à la prédiction *AMIGene* ayant le statut '*newAGC*'. La ligne verticale représente un décalage du cadre de lecture détecté par la *ProFED*. En rose, le nom du décalage de lecture : FS signifie *FrameShift*, 1187150 est la position du décalage du cadre de lecture sur le chromosome et M signifie *Medium* (confiance accordée au décalage du cadre de lecture).

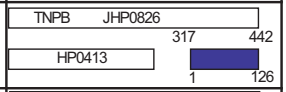
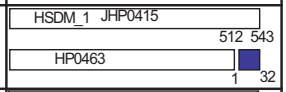
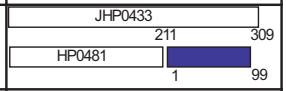
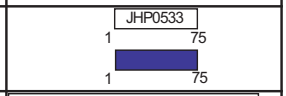
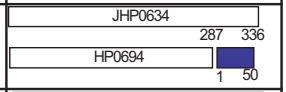
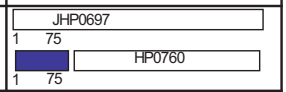
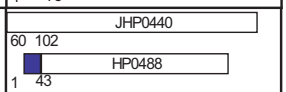
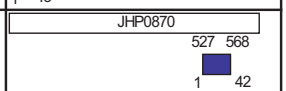
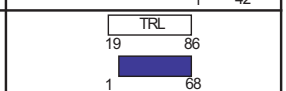
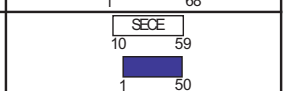
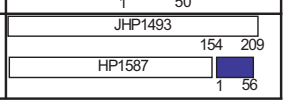
hypothétique identifiée par *GLIMMER* 2.0 dans le fichier des annotations. *GLIMMER* est plus sensible mais moins spécifique que *GeneMark* (Tab. p. 269). Dans certains cas *GLIMMER* peut se tromper, en prédisant par exemple un faux-positif à la place d'un faux-négatif, le faux-positif et le faux-négatif se recouvrant (Fig. 1. *Article IV* p. 316). Après avoir superposé les courbes de *GLIMMER* et de *GeneMark* sur des portions du chromosome de *N. meningitidis* MC58 (serogroup B), je me suis aperçue cette fois-ci que les GNF '*wrongBank*' n'avaient pas des probabilités de codage *GLIMMER* plus élevées que celles des NG '*newAGC*' (Fig. 10.5 B p. 312). Le problème de prédiction de *GLIMMER* 2.0 proviendrait plutôt du calcul du score de codage des CDS et/ou du filtrage des CDS. J'ai donc exécuté *GLIMMER* 2.0 sur le chromosome de *N. meningitidis* MC58 (serogroup B) afin de comprendre d'où venaient ces erreurs.

Dans l'exemple B de la figure 10.5 p. p. 312, les deux CDS, NMB1173.1 '*newAGC*' et NMB1174
'*wrongBank*', sont prédites par *GLIMMER* 2.0 ; de même, les CDS NMB1185 et NMB1185.1 '*newAGC*' sont aussi prédites (les deux GNF '*wrongBank*', NMB1186 et NMB1187, n'ont pas été prédits). De plus, la CDS NMA1345 de *N. meningitidis* Z2491 (Serogroup A) est orthologue à NMB1173.1. En revanche, l'orthologue de NMB1185.1 est aussi une protéine ayant le statut '*newAGC*', NMA1359.1.
Ainsi, l'examen de plusieurs cas de figure nous a permis de recenser au moins deux types de problèmes :

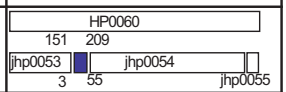
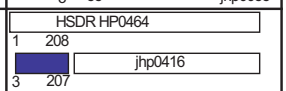
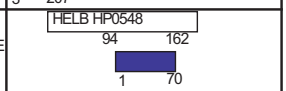
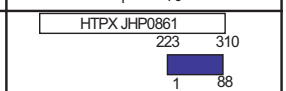
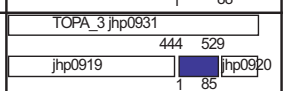
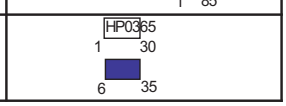
1. Sachant que toutes les CDS prédites par *GLIMMER* 2.0 ont un score de codage supérieur à 0,9, le calcul du score de codage appliqué aux CDS entières (et non pas à une fenêtre glissante) a tendance à favoriser les longues CDS. Par exemple, il est surprenant que le score de codage de NMB1174 soit supérieur à 0,9 vu sa faible prédiction de codage (Fig. 10.5 B p. 312). On a ici à faire à une erreur de prédiction automatique générée par *GLIMMER* 2.0. Même si certaines CDS ont des scores aberrants, quand *GLIMMER* 2.0 sélectionne une CDS '*wrongBank*', il sélectionne aussi la CDS '*newAGC*'. Il arrive aux annotateurs de se tromper en choisissant la CDS '*wrongBank*' au lieu de la CDS '*newAGC*'. En effet, ils ne visualisent pas les courbes de prédictions de codage[6].

2. Des CDS non sélectionnées par *GLIMMER* 2.0 ont été annotées comme ayant été identifiées par *GLIMMER* 2.0. On a alors à faire à une erreur d'annotation qui, en fait, n'a rien à voir avec les prédictions de *GLIMMER* 2.0.

A

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
425513	R	378	Q9ZKV9	442	IS606 TRANSPOSASE (TNPB JHP0826)	8.6e-66	99	HP0412	hypothetical protein (WRONG)	HP0413	transposase-like protein, PS3IS	
481155	R	96	Q9ZM08	543	TYPE I RESTRICTION ENZYME (MODIFICATION SUBUNIT) (HSDM_1 JHP0415)	4.9e-09	90	HP0462	type I restriction enzyme S protein (hsdS)	HP0463	type I restriction enzyme M protein (hsdM)	
505072	D	297	Q9ZLZ2	309	PUTATIVE TYPE II DNA MODIFICATION ENZYME (METHYLTRANSFERASE) (JHP0433)	5.3e-50	96	HP0481	adenine specific DNA methyltransferase (MFOKI)	HP0482	predicted coding region	
615965	R	228	Q9ZLP5	78	PUTATIVE (JHP0533)	3.4e-32	93	HP0585	endonuclease III (nth)	HP0586	predicted coding region	
745418	D	150	Q9ZLE6	336	PUTATIVE OUTER MEMBRANE PROTEIN (JHP0634)	8.4e-20	94	HP0694	predicted coding region	HP0695	hydantoin utilization protein A (hyuA)	
815367	R	279	Q9ZL83	529	PUTATIVE (JHP0697)	1.1e-31	94	HP0760	conserved hypothetical	HP0761	predicted coding region	
932436	R	150	Q9ZLY5	912	PUTATIVE (JHP0440)	1.9e-15	90	HP0881	predicted coding region (SUSPICIOUS)	HP0883	Holliday junction DNA helicase (ruvA)	
996235	D	132	Q9ZKR5	668	PUTATIVE OUTER MEMBRANE PROTEIN (JHP0870)	3.3e-16	97	HP0935	predicted coding region	HP0936	proline/betaine transporter (proP)	
1005895	D	204	O87326	86	TRL Transfert RNA associated Locus *H. pylori* MI 212	3.0e-31	98	HP0945	predicted coding region (WRONG)	tRNA-Leu-1	X	
1280051	R	150	Q9PI37	59	PREPROTEIN TRANSLOCASE SECE SUBUNIT	4.7e-10	56	HP1203	transcription termination factor NusG (nusG)	tRNA-Trp-1	X	
1664725	R	168	Q9ZJ25	209	PUTATIVE (JHP1493)	1.0e-21	91	HP1586	predicted coding region WRONG	HP1587	conserved hypothetical protein	

B

CDS *AMIGene*			DB prot			Match CDSa/DBProt		Previous CDSb		Next CDSb		Alignment DBprot/CDSa (CDSb)
Begin	S	L (bp)	Entry Name	L (aa)	Protein (Gene) name	Proba	Id	/gene	/product (status)	/gene	/product (status)	
61925	D	165	O24900	813	HYPOTHETICAL 93.1 KDA PROTEIN (HP0060)	1.9e-14	74	jhp0053	putative	jhp0054	putative	
448169	R	714	O25211	1055	TYPE I RESTRICTION ENZYME R PROTEIN (HSDR HP0464)	7.7e-97	87	jhp0416	TYPE I RESTRICTION ENZYME (RESTRICTION SUBUNIT) authentic frameshift (hsdR_1)	jhp0417	putative	
547479	D	273	Q9X5H6	185	UREASE-ENHANCING FACTOR probable DNA helicase (HELB CAG-OMEGA HP0548)	1.0e-21	75	jhp0495	cag island protein, cytotoxicity associated IMMUNODOMINANT ANTIGEN	jhp0496	GLUTAMATE RACEMASE (murI)	
957331	D	264	HTPX_HELPJ	310	PROBABLE PROTEASE HTPX HOMOLOG (EC 3.4.24.- JHP0861)	4.7e-42	100	jhp0862	GTP CYCLOHYDROLASE I (folE_1)	jhp0863	GTP CYCLOHYDROLASE I (folE_2)	
1018247	D	261	Q9ZKK4	686	TOPOISOMERASE I (TOPA_3 jhp0931)	1.7e-10	39	jhp0919	topoisomerase I (topA_2)	jhp0920	putative	
1127513	R	105	O25129	30	HYPOTHETICAL 3.2 KDA PROTEIN (HP0365)	4.7e-10	96	jhp1015	putative SUGAR NUCLEOTIDE BIOSYNTHESIS	jhp1016	ribonucleoside-diphosphate reductase I beta chain spore coat polysaccharide biosynthesis	

TAB. 10.8 – CDS '*newAGC*' chez *H. pylori*

A) Souche 26695.

B) Souche J99. Pour la légende, se référer à celle du tableau 10.5 p. 304.

10.3.5 *Article IV* : « The *secE* Gene of *Helicobacter pylori* »

C. Médigue, B. Chun-Yu Wong, M. Chia-Mi Lin, S. Bocs, A. Danchin (2002)
Journal of Bacteriology, 184, 10.

La réannotation des deux souches d'*H. pylori* (26695 et J99) produisent respectivement 0,7 et 0,4% de NG de statut '*newAGC*'. La plupart des NG '*newAGC*' d'*H. pylori* 26695 (neuf sur onze) correspondent à des fragments de CDS qui présentent des similitudes avec des CDS complètes d'*H. pylori* J99 (Tab. 10.8 A p. 315). Des décalages du cadre de lecture seraient présents chez *H. pylori* 26695 et absents chez *H. pylori* J99 (erreur de séquençage ou pseudogène). Nous retrouvons le phénomène inverse chez *H. pylori* J99 : quatre NG '*newAGC*' sur six sont des fragments similaires à des CDS complètes d'*H. pylori* 26695 (Tab. 10.8 B p. 315). Pour les deux souches, la plupart des NG '*newAGC*' sont situés dans des régions de décalage du cadre de lecture détecté par la méthode *ProFED*.

Un NG '*newAGC*' intéressant est détecté chez *H. pylori* 26695. HP0945.1 est localisé à la position 1005895 pb, entre deux gènes d'ARNt spécifiant respectivement tRNA-gly et tRNA-leu. Cette nouvelle CDS est aussi située en face du gène appelé HP0945 qui a été annoté sur le brin inverse et qui a un statut '*wrongBank*'. Nous trouvons une similitude significative de son produit avec un locus associé à un ARN de transfert *trl* d'*H. pylori*. L'expression du gène *trl* d'*H. pylori* a été montré expérimentalement par extension d'amorce sur un pool d'ARNm : ce gène est co-transcrit avec tRNA-gly et révèle une diversité génétique [Dundon *et al.*, 1999]. C'est pour cette raison que nous n'avons pas retrouvé cette CDS dans la souche J99 (cette CDS *trl* est présente chez certaines souches de *pylori* et absente chez d'autres, c'est un marqueur de la diversité génétique intraspécifique). Cette publication démontre aussi expérimentalement qu'il n'y a pas transcription de HP0945, ce qui confirme que cette CDS doit être retirée du jeu d'annotation.

Chez *H. pylori* J99, nous avons trouvé une nouvelle CDS, jhp0495.1, située entre le gène *cagA* (annoté comme antigène de l'îlot de pathogénie *cag*) et le gène *murI* (annoté comme glutamate racémase). Le produit de jhp0495.1 est similaire à Cag-Omega et à HelB d'*H. pylori*. Chez *H. pylori* 26695, le gène HP0548 (annoté comme DNA hélicase putative) révèle un authentique codon de terminaison et ne possède pas de CDS associée. Pourtant, il semble que le fragment en 3′, *helB*, soit fonctionnel et que son produit HelB module l'activité uréase impliquée dans la virulence d'*H. pylori* [McGee *et al.*, 1999].
Chez *H. pylori* J99, il existe trois copies du gène de la topoisomérase I : jhp0108–*topA_1*, jhp0919–*topA_2* et jhp0931–*topA_3*. Nous avons trouvé une nouvelle CDS, jhp0919.1, située entre le gène jhp0919 et le gène jhp0920 (protéine putative). Les trois fragments adjacents, jhp0919–jhp0919.1–jhp0920 correspondent au gène jhp0931. Ce gène a donc été dupliqué dans le génome d'*H. pylori* J99 : la copie *topA_3* est fonctionelle mais pas la copie *topA_2* (pseudogène).
Le NG '*newAGC*' jhp1015.1 (1127513) montre une similitude significative avec HP0365. La deuxième

[6]Dans Imagene, un module *GLIMMER*_curve a été développé, mais dans le package *GLIMMER* 2.0 original, il n'y a pas de *GLIMMER*_curve.

et dernière similitude concerne le gène *trnL* du chloroplaste de l'eucaryote *Chlamydomonas reinhardtii* (Q33366). La jhp1015.1 étant très courte (35 aa), il n'est pas étonnant d'obtenir une E-value non significative (1,8) ; en revanche, il y a 58% d'identité. Il y a 82% de chevauchement entre le tRNA-leu et HP0365, les deux objets étant situés sur le brin direct. Il y a 88% de chevauchement entre tRNA-leu (détecté par tRNAscan mais non annoté) et jhp1015.1, ces deux caractéristiques étant situées sur le brin inverse. Théoriquement, un tel recouvrement ne permet pas l'expression simultanée de l'ARNt et de la CDS ; c'est biologiquement impossible. HP0365 et jhp1015.1 semblent être des artefacts et non des CDS avérées. D'autres arguments vont dans ce sens : il est commun d'observer des artefacts de CDS et des pics de codage parasites recouvrant les ARNt. De plus, aujourd'hui, l'entrée Q33366 a été supprimée car c'est une « fausse » prédiction. Il faut donc se méfier des similitudes entre espèces proches, ou entre souches d'une même espèce, notamment dans le cas des courtes CDS. Si Jhp1015.1 est un artefact, alors HP0365 a le statut '*wrongBank*'.

Enfin, la réannotation des deux souches d'*Helicobacter pylori* (26695 et J99) a permis de mettre en évidence l'« oubli » du gène *secE*. Le label de ce nouveau gène chez *H. pylori* 26695 et *H. pylori* J99 correspond respectivement à HP1203.1 et à jhp1126.1 comme le montre la figure 1 de l'article. Apparemment, ce double « oubli » pourrait être dû à une erreur de prédiction par le programme *GLIMMER*, et/ou à une phase apprentissage de mauvaise qualité. Une validation expérimentale a permis de confirmer la présence de *secE* chez sept isolats d'*H. pylori* et de conclure à l'expression de la translocase SecE, composant essentiel de la principale machinerie de sécrétion.

10.4 Autres génomes réannotés

10.4.1 *M. tuberculosis* H37Rv

Comme nous l'avons déjà vu, le génome de *M. tuberculosis* H37Rv est particulièrement riche en G+C et contient de nombreuses répétitions [Cole *et al.*, 1998]. Cette richesse en G+C se traduit par la présence de très longues ORF chevauchantes tout le long du génome, plusieurs d'entre elles pouvant présenter une probabilité moyenne de codage non nulle. Il est donc tout à fait probable que l'ensemble d'apprentissage utilisé pour calculer la matrice de transition ne soit pas optimal, ce qui expliquerait que certaines régions codantes restent très difficiles à détecter. Aussi, sans résultats de similitude et lorsque plusieurs CDS se chevauchent de façon très importante, comment sélectionner la CDS à conserver ? De nombreuses CDS annotées par les auteurs ont en effet une probabilité de codage inférieure à 0,15 (nous trouvons même des valeurs à 0,01), et sont décrites comme étant des protéines hypothétiques. Elles recouvrent généralement des CDS uniques à *AMIGene*, qui sont plus petites mais qui ont une meilleure probabilité moyenne de codage (Tab. 10.1 B p. 290). D'autres cas particuliers, délicats à traiter dans le cas d'une procédure automatique de sélection de phases codantes, se rencontrent chez *M. tuberculosis* H37Rv et chez la plupart des organismes pathogènes dont le génome contient des éléments répétés, impliqués dans les processus de variabilité antigénique. S. Cole, le coordinateur du consortium d'annotation de *M. tuberculosis* H37Rv [Cole *et al.*, 1998], a soigneusement analysé les résultats de réannotation obtenus pour cette bactérie.

Parmi les *potential New Gene* (NG) ayant un statut '*newAGC*' (54 gènes, soit 1,32%), 61,5% ont été intégrés dans une version mise à jour des annotations, 31,6% avait été identifié précédemment mais n'ont pas été incorporés dans le jeu d'annotation INSD, et 7% ont été considérées insuffisamment convaincante (Tab. 10.6 p. 320 [Camus *et al.*, 2002]).
Parmi les 61 CDS présentées dans ce tableau, 29 mettent en évidence des points de mutation et 4 ont permis au contraire de corriger la séquence et de fusionner des CDS (disparition des décalages du cadre de lecture). Certains des NG '*newAGC*' sont répertoriées dans la banque Swiss-Prot sous l'étiquette *NOT-ANNOTATED-CDS* (par exemple, une protéine régulatrice de la famille AraC en position 1571045 pb), alors que d'autres présentent de fortes similitudes avec des protéines d'autres organismes (le gène *carB* de *S. coelicolor* pour une CDS de plus de 3000 pb localisée en 1557099 bp, ou encore un gène de transporteur d'*Arthrobacter sp.* pour une CDS de 1302 bp localisée en 61303 bp).
Si ces séquences ne sont actuellement pas répertoriées dans les banques de protéines pour la bactérie *M. tuberculosis* H37Rv, une interrogation de la base de données TubercuList a toutefois permis d'en retrouver (31,6%). Ces observations montrent qu'entre les données d'annotations répertoriées dans les banques de séquences et celles qui sont gérées dans les différentes bases de données spécialisées, les différences peuvent être parfois très importantes. En fait, un tiers des NG '*newAGC*' avaient été originellement détectés mais mal annotés en utilisant des *feature keys* inappropriées telles que *misc_feature* ou *mRNA*, au lieu de *CDS*.
Certains NG '*newAGC*' et *annotated Gene Not Found* (GNF) '*wrongBank*' se sont mutuelle-

Bk_label	close to gene	St	begin	end	L	Pc	O	Prot_id	description	Sc	Evalue	id%	comment	MYCTC
Rv0157A	Rv0157/RV0158	R	186495	186623	129	0,57	N	O53976	23.6 KDA HP	173	3,00E-06	63	gene fragment, SIMTO Cterm of O53976 -> FS	
Rv0164	Rv0163/RV0165c	D	193626	194111	486	0,82	N	O06090	HP ML2629	594	6,13E-61	76	change mRNA feature key to CDS	MT0173
Rv0281A	Rv0281/Rv0282	R	341986	342291	306	0,57	N		NO SIMILARITY				S. Cole: no significant, SIMTO *S. natalensis* , Pc=0.63 mat III, in front of Rv0282 -> start problems ?, overlaps UNIQUE_AMIGene -> FS ?	MT0294
Rv0392A	Rv0392c/Rv0393	D	472914	473105	192	0,70	N	O33360	HP RV0515.	223	1,06E-17	50	included in Rv0393 -> compensated FS but in fact S. cole did not add /pseudo in Rv0393	
Rv0470A	Rv0470c/Rv0471c	R	561854	562294	441	0,42	N	Q9KNQ8	1,4-DIHYDROXY-2-NAPHTHOATE OCTAPRENYLTRANSFERASE.	83	0,051	30	overlaps a UNIQUE_AMIGENE and Rv0471c -> FS	MT0487
Rv0492A	Rv0492c/Rv0493c	R	583375	583704	330	0,67	N		NO SIMILARITY					
Rv0500A	Rv0500/Rv0501	D	591111	591347	237	0,55	N	Q49828	B2168_C1_172 (DNA-BINDING PROTEIN).	339	1,43E-31	83	already added	MT0521
Rv0521	Rv0521c	D	612598	612903	306	0,49	Y	Q9KJ21	SARCOSINE-DIMETHYLGLYCINE METHYLTRANSFERASE.	109	1,40E-05	41	in front of W Rv0521c (removed)	MT0543
Rv0522	Rv0521/Rv0523c	D	613038	614342	1305	0,70	N	O66184	TRANSPORTER.	1221	7,14E-133	54	already added (missing by accident)	MT0544
Rv0590A	Rv0590/Rv0591	D	688808	689062	255	0,62	N	O07414	37.7 KDA HP (VIRULENCE FACTOR MCE FAMILY PROTEIN).	283	4,35E-25	62	SIMTO Cterm of other Mce, overlaps Rv0590 -> FS	MT0620
Rv0634A	Rv0634c/Rv0635	D	731113	731364	252	0,32	N	Q10848	HP RV2009.	97	0,001	39	UNIQUE_AMIGene already added	MT0662.1
Rv0634B	Rv0634c/Rv0635	D	731712	731879	168	0,54	N	P96925	50S RIBOSOMAL PROTEIN L33 TYPE 2.	305	1,36E-27	100	already added	MT0663
Rv0724A	Rv0724/Rv0725c	R	817531	817866	336	0,81	N	P71987	HP RV1729C.	306	8,37E-28	59	SIMTO Cterm of RV1729C, overlaps Rv0725c -> FS	
Rv0749A	Rv0749/Rv0750	R	841737	841874	138	0,61	N	P71644	42.8 KDA HP.	101	0,0002	46	gene fragment, SIMTO part of Rv2807 -> FS	MT0773.1
Rv0755A	Rv0755c/Rv0756c	R	850342	850527	186	0,40	N	Q9RI80	PUTATIVE NONCOMPOSITE TRANSPOSON TRANSPOSASE.	111	4,17E-05	50	gene fragment -> FS	MT0780
Rv0787A	Rv0787/Rv0788	D	882524	882763	240	0,58	N	O05755	8.6 KDA HP MLCB5.24.	290	2,00E-26	75	already added	MT0812
Rv0946A	Rv0946c/Rv0947c	R	1056786	1056998	213	0,44	N	P17944	ANTIGEN 85-A PRECURSOR (32 KDA)	115	4,00E-06	66	S. cole: no significant	MT0973
Rv1000c	Rv1000	D	1116531	1117148	618	0,50	Y	Q9K4L2	HP SCO7302.	599	2,87E-61	57	in front of W Rv1000 (removed)	MT1029
Rv1116A	Rv1116/Rv1117	R	1241115	1241390	276	0,40	N	P94981	30.2 KDA HP.	234	9,00E-20	68	gene fragment, SIMTO Cterm of Rv1646 -> FS	MT1148
Rv1322A	Rv1322/Rv1323	R	1485313	1485771	459	0,70	N	Q49717	B1549_F2_87.	551	0	77		MT1364
Rv1384	*carA* /*pyrF*	D	1555971	1557101	1131	0,94	N	P14846	CARBAMOYL-PHOSPHATE SYNTHASE LARGE CHAIN (EC 6.3.5	2689	0	51	already added (missing by accident)	
Rv1395	Rv1394c/Rv1396c	D	1571047	1572081	1035	0,71	N	P71663	HYPOTHETICAL TR RV1395.	1666	0	95	change mRNA feature key to CDS	MT1440
Rv1473A	Rv1473/Rv1474c	D	1662381	1662572	192	0,68	N	Q9RJ29	PUTATIVE TR.	129	1,00E-07	53		MT1520
Rv1489	Rv1489c	D	1678552	1678908	357	0,33	Y	Q9K543	Putative invasion protein [Fragment]	86	0,026293	34	UNIQUE_AMIGENE (Pc=0,807 mat III) in front of Rv1489c (removed), fragment?	MT1534
Rv1489A	Rv1489/Rv1490	D	1678942	1679172	231	0,53	N	P71774	PROBABLE METHYLMALONYL-COA MUTASE ALPHA-SUBUNIT (E	138	8,00E-09	51	gene fragment, SIMTO part of alpha subunit of many methylmalonyl-CoA mutases -> FS	MT1535
Rv1498A	Rv1498c/Rv1499	R	1690134	1690346	213	0,56	N	Q9RCZ5	9.6 KDA HP.	257	9,00E-23	65	already added, found by proteomics	MT1547
Rv1575A	Rv1575/Rv1576c	D	1780456	1780653	198	0,53	N	O06616	11.4 KDA HP.	191	5,00E-15	75	SIMTO Nterm of RV1574, included in Rv1575 -> FS	MT3573.13
Rv1638A	Rv1638/Rv1639c	R	1846716	1846973	258	0,48	N	Q49845	B2235_C3_214.	96	0,001	61	gene fragment, SIMTO Cterm of 35kd immunogenic protein -> FS	MT1676
Rv1706A	PPE23/ Rv1707	R	1934482	1934649	168	0,54	N	P71836	56.6 KDA HP CY369.27C.	94	0,001	64	gene fragment, SIMTO part of export proteins -> FS	MT1747
Rv1765A	Rv1765c/Rv1766	R	1999142	1999357	216	0,42	N	Q54336	ORF2 OF THE IS3 FAMILY.	185	3,00E-14	55	gene fragment, SIMTO part of many transposase genes including IS6110 -> FS	MT1803
Rv1792	PE19/ Rv1793	D	2030347	2030523	177	0,52	N	P96363	PUTATIVE ESAT-6 LIKE PROTEIN 2.	302	5,00E-28	100	change mRNA feature key to CDS, in-frame stop codon (Nterm fragment), SIMTO QILSS family (Rv1038c)-> FS	MT1067
Rv1792	PE19/ Rv1793	D	2030524	2030643	120	0,71	N	P96363	PUTATIVE ESAT-6 LIKE PROTEIN 2.	196	6,26E-15	100	change mRNA feature key to CDS, in-frame stop codon (Cterm fragment noStatusAGC), SIMTO QILSS family (Rv1038c)-> FS	MT1067

FIG. 10.6 – CDS '*newAGC*' chez *M. tuberculosis* H37Rv (pour la légende voir TAB. 10.9 p. 324)

Bk_label	close to gene	St	**begin**	end	**L**	Pc	O	Prot_id	description	Sc	Evalue	id%	comment	MYCTC
Rv1931c	Rv1931c	R	2182687	2183187	501	0,67	N	Q9RJG8	ARAC FAMILY TR.	202	5,00E-16	54	correction of sequencing error leads to the fusion of Rv1931A and Rv1931c	
Rv1946A	Rv1946c	D	2197613	2197849	237	0,56	Y		NO SIMILARITY				S. cole: no significant, A included in W Rv1946c (kept)	
Rv2024A	Rv2024c/Rv2025c	D	2270504	2271076	573	0,51	N		NO SIMILARITY				S. cole: no significant, SIMTO very HP Rv2086	
Rv2063	Rv2063c	D	2320831	2321064	234	0,44	Y		NO SIMILARITY				A in front of W Rv2063c (removed)	MT2122
Rv2160A	Rv2160c/Rv2161c	R	2421643	2422278	636	0,42	N	Q9JN89	24.1 kDa HP Putative lactone-dependent TR (TetR-family), MmfR	113	5,00E-05	40	SIMTO Nterm of tetR-family TR, includes Rv2160 -> FS	
Rv2219A	Rv2219/Rv2220	R	2486994	2487416	423	0,80	N	Q9S2N9	16.6 KDA HP.	148	2,00E-09	35	Probable integral membrane protein	MT2277
Rv2250A	Rv2250c/Rv2251	D	2525402	2525821	420	0,60	N	Q9RJ97	PUTATIVE FLAVOPROTEIN.	163	5,00E-11	38	SIMTO Nterm of flavoprotein, overlaps Rv2251 -> FS	MT2311
Rv2282A	Rv2283	R	2555925	2556251	327	0,55	Y		NO SIMILAR				S. cole: no significant, included in W Rv2283 (kept)	MT2341
Rv2306A	Rv2306c	D	2577108	2577701	594	0,40	Y	P96915	25.2 KDA HP RV0625C.	307	1,92E-27	53	in front of W Rv2063c (removed), UNIQUE_AMIGENE overlaps Rv2306B -> FS	MT0653
Rv2306B	Rv2306c	D	2577488	2577922	435	0,45	Y	P96915	25.2 KDA HP RV0625C.	308	4,00E-28	78	in front of W Rv2063c (removed), overlaps Rv2306A -> FS	MT0653
Rv2401A	Rv2401	R	2698042	2698245	204	0,53	Y	Q49760	10.3 KDA HP B1937_F2_47.	169	3,00E-12	50	in front of W Rv2401 (kept, short overlap); conserved membrane protein	MT2473
Rv2427A	Rv2427c/Rv2428	R	2725595	2725861	267	0,45	N	P52677	Probable hydrogen peroxide-inducible genes activator	171	2,00E-12	72	change misc_feature key to CDS but in fact S. cole did not, /pseudo -> FS	
Rv2631A	Rv2631/Rv2632c	D	2958364	2958870	507	0,61	N	O59245	98.1 KDA HP PH1602.	440	2,00E-43	54	correction of sequencing error leads to the fusion of Rv2631 and Rv2631A	MT2707
Rv2702A	*ppgK*/*sigA*	R	3017738	3018124	387	0,42	N		NO SIMILARITY				S. cole: no significant	
Rv2737A	Rv2737c/Rv2738c	D	3051619	3051792	174	0,41	N	Q9X9U6	18.7 KDA HP (FRAGMENT).	153	2,00E-10	59	gene fragment, SIMTO central part of cysteine-rich *glgA* -> FS	
Rv2803	Rv2803c	D	3111822	3112289	468	0,56	Y	O05910	17.4 KDA HP.	123	2,00E-06	42	in front of W Rv2803c (removed), SIMTO Cterm of Rv0918	MT2871
Rv2856A	Rv2856/Rv2857c	D	3167868	3168200	333	0,72	N	O66901	Hydrogenase expression/formation protein B	191	1,00E-14	44	change misc_feature key to CDS but in fact S. cole did not, Nterm fragment of *hypB* -> FS	
Rv2856B	Rv2856/Rv2857c	D	3168251	3168427	177	0,62	N	Q43949	Hydrogenase expression/formation protein B	190	7,00E-15	52	change misc_feature key to CDS but in fact S. cole did not, Cterm fragment of *hypB* -> FS	
Rv3022A	*PPE48*/Rv3023c	R	3380679	3380993	315	0,62	N	O53690	PE-FAMILY PROTEIN.	165	1,00E-11	59	PE29, best predicted by mat II (leading)	MT3106.1
Rv3128c	Rv3127/Rv3129	R	3493168	3493572	405	0,77	N	P71644	42.8 KDA HP.	226	1,00E-18	47	change mRNA feature key to CDS, in-frame stop codon (Cterm fragment), SIMTO RV2807-> FS	MT2874
Rv3128c	Rv3127/Rv3129	R	3493600	3494262	663	0,70	N	P71644	42.8 KDA HP.	411	6,00E-40	46	change mRNA feature key to CDS, in-frame stop codon (Nterm fragment), SIMTO RV2807-> FS	MT2874
Rv3221A	*TB7.3*/Rv3222c	R	3598051	3598356	306	0,42	N	Q9XCD7	11.6 KDA HP.	153	2,00E-10	81	correction of sequencing error leads to the fusion of Rv3221A and Rv3221B, POSSIBLE ANTI-SIGMA FACTOR	MT3318
Rv3224A	Rv3224/Rv3225c	D	3600635	3600823	189	0,42	N	Q9XCD9	19.1 KDA HP.	95	0,001	62	gene fragment, SIMTO central part of ML0799 -> FS	
Rv3224B	Rv3224/Rv3225c	D	3600801	3601019	219	0,29	N	Q9XCD9	19.1 KDA HP.	154	4,14E-10	68	UNIQUE_AMIGene fragment, SIMTO Cterm of ML0799 -> FS	
Rv3395A	Rv3395c/Rv3396c	D	3811719	3812345	627	0,45	N		NO SIMILARITY				probable membrane protein, SIMTO lipoprotein	MT3503
Rv3566c	*nat*	R	4007331	4008182	852	0,64	N	O86309	ARYLAMINE N-ACETYLTRANSFERASE (EC 2.3.1.5).	333	3,00E-31	59	correction of sequencing error leads to the fusion of Rv3565A and Rv3666c, FS ? inactive or decrease *nat* expression -> isoniazid resistance	MT3671
Rv3678A	Rv3678c/Rv3679	R	4118530	4118691	162	0,45	N	Q9XA37	6.1 KDA HP.	199	6,00E-16	78		MT3780
Rv3770A	Rv3770c/3771c	R	4215881	4216063	183	0,49	N	P71639	52.4 KDA HP.	159	3,00E-11	80	Transposase fragment, SIMTO part of Rv2812 -> FS, continuation of Rv3770B	MT3878
Rv3770B	Rv3770c/3771c	R	4216078	4216269	192	0,42	N	P71639	52.4 KDA HP.	293	6,00E-27	93	Transposase fragment, SIMTO Nterm of Rv2812 -> FS	MT3879

FIG. 10.6 – CDS *'newAGC'* chez *M. tuberculosis* H37Rv

ment confirmés (dans huit cas, un NG '*newAGC*' recouvre un GNF '*wrongBank*' parmi 25 GNF '*wrongBank*'). Plus de 50% des NG '*newAGC*' sont impliqués dans des décalages du cadre de lecture probables de la séquence génomique. Les produits des NG '*newAGC*' sont similaires à des protéines hypothétiques conservées (43%), à des séquences d'insertion ou des protéines de prophage (23%) ou à des protéines d'autres fonctions biologiques (34%) comme des protéines régulatrices, d'export ou des lipoprotéines. Nous avons aussi analysé le *potential New Gene* (NG) situé sur le brin direct à la position 1678550 pb entre Rv1488 (protéine hypothétique) et Rv1490 (protéine de membrane probable). Ce NG est sans statut car il recouvre le GNF Rv1489c (Pc=0,21). Son produit (de 118 aa) est similaire à une protéine d'invasion de *Mycobacterium paratuberculosis*. Ce NG sans statut remplace indubitablement l'entrée précédente Rv1489c qui avait été annotée au même endroit, mais sur le brin inverse. Ce cas de réannotation décrit une fois de plus les problèmes rencontrés avec les ombres du codant dans les CDS contenant des répétitions.

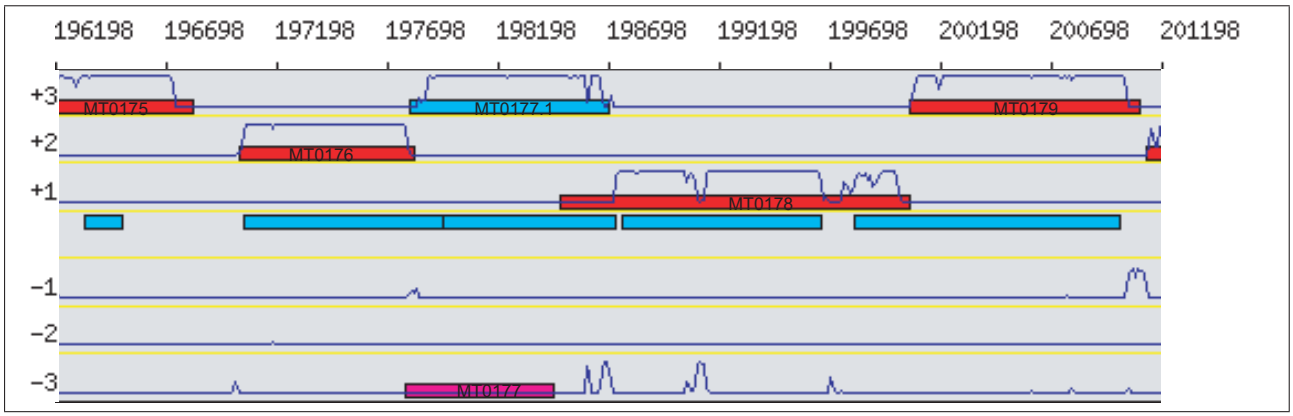


FIG. 10.7 – Exemple de CDS '*newAGC*' et '*wrongBank*' chez *M. tuberculosis* CDC1551
Cette région présente des répétitions comme le montre rectangles cyan de la phase 0 (répétitions longues de type *distant* prédites par le programme Nosferatu ; [Achaz *et al.*, 2002]). MT0177 est une protéine hypothétique identifiée par *GLIMMER* 2.0 ; elle a le statut '*noStatusBank*' dans le cas des résultats de MYCTC** et '*wrongBank*' dans le cas des résultats de MYCTC*** (TAB. 10.2 p. 295). MT0177.1 est une CDS prédite par *AMIGene* ; elle a le statut '*ambiguousAGC*' dans le cas des résultats de MYCTC** et '*newAGC*' dans le cas des résultats de MYCTC***.

Nous avons intégré et réannoté le jeu d'annotations de référence de *M. tuberculosis* H37Rv et le jeu d'annotations *RefSeq* de *M. tuberculosis* CDC1551 (NC_002755) dans PkGDB (voir p. 213 et p. 251). Dans le tableau 10.2 p. 295, nous présentons quatre résultats de réannotation automatique pour *M. tuberculosis* H37Rv et trois pour *M. tuberculosis* CDC1551 :

1. l'« ancienne » façon de réannoter avec l'ancien jeu d'annotation (ligne MYCTU*),
2. la façon « actuelle » de réannoter avec les jeux actuels avec trois matrices de transition (lignes MYCTU et MYCTC),
3. ces mêmes conditions mais après avoir recommencé la validation automatique des paramètres d'*AMIGene* avec quatre matrices (lignes MYCTU** et MYCTC**) et
4. ces mêmes conditions mais en utilisant les filtres Seg et Xnu pour masquer respectivement les régions de basse complexité et celles contenant des répétitions de courte périodicité dans les sé-

quences protéiques au moment de la recherche de similitude avec BioFacet (lignes MYCTU*** et MYCTC***).

Nous observons une augmentation de la proportion de CDS communes entre les résultats MYCTU* et MYCTU (CG : 95,73 à 98,75%) reflétant plus une augmentation de CDS prédites (AP : 4096 à 4739) qu'une amélioration de la prédiction due à l'utilisation de matrices adaptées à l'usage des codons synonymes. En revanche, les résultats MYCTU** démontrent que la partition des CDS en quatre classes est plus adaptée que celle en trois classes dans le cas de MYCTU (AP : 4739 à 4387 et CG : 98,75 à 99,07%).
La progression est similaire lorsque l'on compare les résultats MYCTC et MYCTC** (AP : 4736 à 4448 et CG : 93,66 à 94,43%) mais la proportion de CDS communes est inférieure à celle de MYCTU (CG : 98,75 à 99,07%). Sachant que les deux chromosomes montrent plus de 99% d'identité au niveau nucléotidique [Delcher *et al.*, 1999a], il est impossible d'avoir une différence de densité de gènes aussi importante (plus de 7% ; Tab. p. 253). Cette hétérogénéité d'annotation entre les deux jeux de CDS doit encore être plus prononcée si l'on regarde le nombre de CDS différentes (*i.e.* les CDS de *M. tuberculosis* CDC1551 qui n'ont pas d'orthologues chez *M. tuberculosis* H37Rv ; voir p. 264).
Enfin, si l'on compare les résultats MYCTU** et MYCTU***, on observe uniquement des différences au niveau des nombres des CDS '*ambiguousAGC*' et '*newAGC*' : quatre CDS '*newAGC*' sont devenues '*ambiguousAGC*'. Si les séquences traduites de ces quatre CDS sont masquées, elle ne présentent alors plus de similitudes avec les séquences protéiques de la SWALL et deviennent '*ambiguousAGC*'. Si l'on compare les résultats MYCTC** et MYCTC***, on observe des différences à la fois au niveau des nombres de CDS '*ambiguousAGC*' et '*newAGC*' et au niveau des nombres de CDS '*suspiciousBank*', '*wrongBank*' et '*noStatusBank*'. L'exemple de la figure 10.7 p. 322 montre que l'on masque la séquence traduite de MT0177, elle ne présente alors plus de similitudes avec les séquence de la SWALL : elle passe donc du statut '*noStatusBank*' au statut '*wrongBank*' ; ce qui a pour conséquence de faire passer MT0177.1 du statut '*ambiguousAGC*' au statut '*newAGC*'. Ainsi dans le cas de MYCTU*** et MYCTC***, l'utilisation de filtres pour masquer les séquences protéiques au moment de la recherche de similitude améliore les résultats d'attribution de statuts de réannotation.

10.4.2 *B. subtilis*

Dans le tableau 10.2 p. 295, nous présentons deux résultats de réannotation du même jeu de référence de *B. subtilis* (voir p. 220), issus des deux versions du processus de réannotation semi-automatique. Les résultats de la ligne BACSU* correspondent à l'« ancienne » façon de réannoter (une seule matrice, paramètres uniformes dont la valeur a été ajusté de manière empirique, etc.) et ceux de la ligne BACSU correspondent à la manière dont nous réannotons « actuellement » les chromosomes procaryotes. Nous observons que cette dernière a permis d'augmenter la proportion de CDS communes (96,37 à 99%) et de diminuer le nombre de CDS uniques aux banques (149

Bk_label	close to gene	St	begin	end	L	Pc	O	Prot_id	description	Sc	Evalue	id%	comment
Bs3568.1	*ggaA*	D	3671959	3672555	597	0,86	N	AAK33308	Putative UDP-glucose pyrophosphorylase	474	0	48	teichoique operon
Bs2122.1	*yomU/T*	D	2240452	2240961	510	0,49	N	O64057	HYPOTHETICAL 21.7 KDA PROTEIN.	776	0	92	phage operon
Bs1682.1	*ymfD/E*	D	1755258	1755674	417	0,53	N	Q9K7Q2	MULTIDRUG RESISTANCE PROTEIN.	357	3,00E-33	58	FS
Bs0608.1	*ydiQ/ydiR*	R	657668	658054	387	0,80	Y	YHXB_BACSU	PROBABLE PHOSPHOMANNOMUTASE (EC 5.4.2.8) (PMM).	387	4,10E-37	68	FS, in front of W *ydiQ*
Bs2922.1	*dnaE*	D	2993786	2994136	351	0,82	N	Q9K837	BH3170 PROTEIN.	196	8,00E-15	38	serine protease?
Bs1107.1	*yitP/Q*	D	1183961	1184308	348	0,58	N	Q9RD92	CONSERVED HYPOTHETICAL PROTEIN SCE20.33C.	196	8,00E-15	40	FS, HTH_ MarR domain, sulfur?
Bs3633.1	*ywpE/D*	R	3741137	3741481	345	0,53	N	Q9CJ55	HYPOTHETICAL PROTEIN YBEF.	158	2,00E-10	49	SIMTO *L. lactis* , *B. halodurans*
Bs3126.1	*tgl*	R	3212302	3212643	342	0,46	N		NO SIMILARITY				serine phosphatase?
Bs2359.1	*yqxK/nudF*	R	2457363	2457701	339	0,78	N	Q9KCP4	BH1525 PROTEIN.	279	2,00E-24	55	SIMTO *B. pumilus*
Bs0435.1	*ydaQ*	R	490332	490655	324	0,55	Y		NO SIMILARITY				in front of W *ydaQ*
Bs0306.1	*lctP/mdr*	R	331995	332309	315	0,78	N	P96712	MULTIDRUG TRANSPORTER.	374	1,00E-35	80	FS with *mdr*
Bs0333.1	*nasA*	D	363398	363706	309	0,58	Y	Q9X4K3	NITRATE TRANSPORTER.	113	3,00E-05	45	compensated FS with *nasA*
Bs2962.1	*yttP/ytsP*	D	3032210	3032512	303	0,68	N	Q9K7Z9	BH3208 PROTEIN.	304	2,00E-27	68	
Bs3800.1	*ywdC*	R	3899538	3899837	300	0,76	Y	Q9RT15	CONSERVED HYPOTHETICAL PROTEIN.	190	3,00E-14	38	in front of W *ywdC* , SIMTO PadR_TR family
Bs1838.1	*trnSL-ArgII/yoeC*	R	2002610	2002894	285	0,58	N	Q9K5P9	BH4039 PROTEIN.	187	6,00E-14	48	phage integrase
Bs1947.1	*yojF/E*	R	2121871	2122149	279	0,71	N	O68260	YojE.	466	1,10E-43	100	*yojE->yojD* , 2 *yojE* in TrEMBL
Bs3265.1	*yurT/U*	R	3354072	3354350	279	0,51	N	Q9KB45	TRANSCRIPTIONAL REGULATOR (ARSR FAMILY).	129	3,00E-07	35	HTH_AsR domain
Bs1296.1	*dppE/ykfA*	D	1365195	1365437	243	0,62	N	Q9HLJ4	HYPOTHETICAL PROTEIN TA0234.	104	0.0002	37	FS
Bs2190.1	*metA/ugtP*	D	2305241	2305480	240	0,69	N	P54167	HOMOSERINE O-SUCCINYLTRANSFERASE (EC 2.3.1.46)	439	3,00E-43	100	FS corrected in Swiss-Prot
Bs1691.1	*ymfM/pgsA*	D	1761621	1761845	225	0,52	N	P94510	HYPOTHETICAL 34.7 KDA PROTEIN.	360	3,00E-34	97	FS (Cterm) with *ymfM* , complete in TrEMBL
Bs3284.1	*yusM/N*	D	3372794	3373015	222	0,66	N	Q9K781	BH3489 PROTEIN.	222	3,00E-18	58	SIMTO *B. halodurans* (short)
Bs1823.1	*yngG/H*	R	1952154	1952375	222	0,71	N	Q9R9I3	YNGXX.	335	2,00E-31	93	FS with *YngG* , biotin lipoyl binding domain
Bs0889.1	*ygaO/trnSL-GlyI*	R	966008	966229	222	0,64	N	Q9V101	REPRESSOR PROTEIN, PUTATIVE.	186	4,00E-14	53	HTH_Xre domain
Bs1129.1	*yjaV*	D	1205625	1205846	222	0,58	N	O32435	HYPOTHETICAL PROTEIN.	303	1,00E-27	78	FS (Cterm) overlaps *yjaV* , complete in TrEMBL
Bs0589.1	*ydhU/trnE-Arg*	R	634129	634335	207	0,45	N	Q9KAU6	MANGANESE-CONTAINING CATALASE.	109	4,00E-05	47	FS (Nterm)
Bs2771.1	*queA/ruvB*	R	2834183	2834383	201	0,48	N	Q9KDI7	BH1226 PROTEIN.	108	5,00E-05	47	TM domain? quorum sensing Sec. Syst?
Bs2699.1	*yraB/adhA*	D	2755212	2755406	195	0,48	N	Q9CFE3	HYPOTHETICAL PROTEIN YPHJ.	197	2,00E-15	71	
Bs2923.1	*ytrI/ytqI*	D	2994738	2994929	192	0,42	N	Q9K835	BH3172 PROTEIN.	116	6,00E-06	35	SIMTO *B. halodurans* (short)
Bs2745.1	*glnP/yrrD*	R	2804728	2804919	192	0,49	N	Q9KDE8	BH1265 PROTEIN.	248	3,00E-21	70	SIMTO *B. halodurans* (short)
Bs0256.1	*ycbN*	D	279618	279797	180	0,61	N	Q9KEN4	BACITRACIN ABC TRANSPORTER (ATP-BINDING PROTEIN)	185	6,00E-14	65	FS (Cterm) overlaps *ycbN*
Bs1381.1	*ykvS*	D	1447345	1447515	171	0,58	N	Q9KAG0	BH2327 PROTEIN.	137	2,00E-08	50	in front of *ykvS* , SIMTO *B. halodurans* (short)
Bs3194.1	*yukJ*	R	3279319	3279489	171	0,68	N	Q9Z388	PUTATIVE SMALL CONSERVED HYPOTHETICAL PROTEIN.	198	2,00E-15	64	SIMTO Gram + rich G+C, iron?
Bs0356.1	*ycxD/sfp*	R	407032	407199	168	0,70	N	P39135	4'-PHOSPHOPANTETHEINYL TRANSFERASE (EC 2.-.-.-)	299	4,00E-27	100	FS in strain 168, other subtilis strains ?
Bs1999.1	*yosV/U*	R	2157318	2157485	168	0,47	N	O30602	HYPOTHETICAL 6.6 KDA PROTEIN.	221	4,00E-18	78	already annotated as short protein in TrEMBL
Bs1968.1	*yozE/kamA*	R	2138378	2138539	162	0,57	N	O30470	YOKU.	254	6,00E-22	98	FS (Cterm), complete in TrEMBL
Bs0608.2	*ydiQ/ydiR*	R	658361	658519	159	0,45	N	YHXB_BACSU	PROBABLE PHOSPHOMANNOMUTASE (EC 5.4.2.8) (PMM).	139	1,00E-08	88	FS
Bs0070.1	*yacB*	D	79713	79865	153	0,47	N	Q9F985	PUTATIVE 32 KDA REPLICATION PROTEIN.	154	2,00E-10	76	FS (Cterm) overlaps *yacB* , SIMTO *B. halodurans*
Bs1722.1	*pksS*	R	1857760	1857909	150	0,45	N	P33271	CYTOCHROME P450 107B1 (EC 1.14.-.-) (P450CVIIB1).	123	9,00E-07	51	FS (Cterm) overlaps *pksS* , SIMTO
Bs3270.1	*yurY/Z*	R	3360001	3360138	138	0,71	N	Q9K797	BH3472 PROTEIN.	109	4,00E-05	48	FS: short in *B. halodurans,* long in *L. lactis*
Bs3901.1	*yxjA/ysiT*	R	4006005	4006271	267	0,80	N		NO SIMILARITY				SIMTO *M. jannaschii*
Bs1183.1	*yjcE*	D	1255421	1255675	255	0,52	Y		NO SIMILARITY				in front of W *yjcE*
Bs1743.1	*ynbA/B*	D	1875513	1875761	249	0,72	N		NO SIMILARITY				magnesium pemease Neisseria?
Bs4056.1	*yybO/N*	R	4170420	4170659	240	0,75	N		NO SIMILARITY				SIMTO
Bs0653.1	*purD/yezC*	R	710794	711030	237	0,55	N		NO SIMILARITY				FS (Cterm) overlaps *yerC* , SIMTO AsnC_R
Bs1264.1	*xkdJ/K*	D	1331497	1331715	219	0,53	N		NO SIMILARITY				SIMTO *L. innocua*
Bs1179.1	*yjcB*	D	1252127	1252333	207	0,69	Y		NO SIMILARITY				FS, overlaps W *yjcB*
Bs3911.1	*yxiL*	R	4017804	4017992	189	0,47	Y		NO SIMILARITY				FS? overlaps W *yxiL* , SIMTO *C. muridarum*
Bs3912.1	*yxiL/K*	R	4018023	4018154	132	0,46	Y		NO SIMILARITY				FS? overlaps W *yxiL*

TAB. 10.9 – Nouvelles CDS chez *B. subtilis*

Abreviations : Bk_label, bank label ; S, St, strand ; L, length (bp) ; Pc, coding average probability ; O, overlap between a unique bank CDS and a unique *AMIGene* CDS (Y, yes; N, no) ; Prot_id, Entry name in SWALL ; Sc, Blast2p alignment score ; Id%, identity percentage ; FS, frameshift ; SIMTO, similar to ; A, *'ambiguousAGC'* ; N, *'newAGC'* ; W, *'wrongBank'* ; HP, hypothetical protein ; HTH, helix turn helix ; TM, transmembrane ; TR, transcriptional regulator ; Sec. Syst., secretion system ; Cterm, C-terminus, Nterm, N-terminus ; M, codon usage matrix number ; ϕ, phage. Voici la correspondance entre les couleurs des CDS et leur statut de réannotation : *'common'* (rouge), *'newAGC'* (bleue), *'ambiguousAGC'* (cyan), *'noStatusAGC'* (vert), *'uniqAGC'* (blanc), *'wrongBank'* (jaune), *'suspiciousBank'* (orange), *'noStatusBank'* (rose) et *'uniqBank'* (violet). Les résultats de similitude ont été triés selon leur E-value (le résultat présenté correspond à la meilleure E-value).

à 41), de CDS '*wrongBank*' (29 à 10) et '*suspiciousBank*' (26 à 15). Bien que le nombre de CDS prédites par *AMIGene* ait augmenté (4193 à 4586), le nombre de '*newAGC*' reste du même ordre de grandeur (50 à 47). Le tableau 10.9 p. 324 présente une sélection de *potential New Gene* intéressants

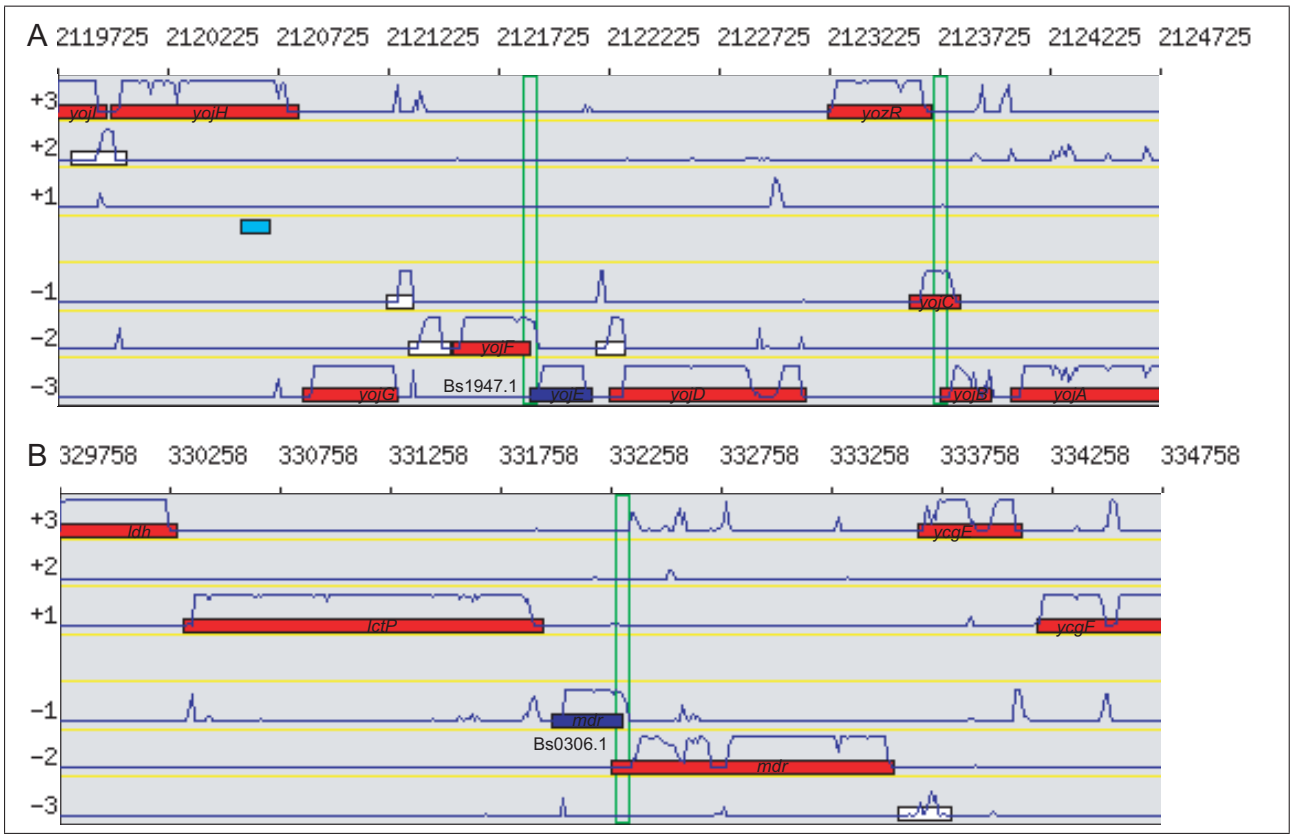


FIG. 10.8 – Exemple de '*newAGC*' chez *B. subtilis*
A) *yojD* et *yojE* chez *B. subtilis*. Dans les annotations actuelles de *B. subtilis* (R16.1), *yojD* correspond à *yojE* et *yojE* n'est pas annotée. Le décalage du cadre de lecture concerne uniquement *yojF* (démarrage tardif de la CDS relativement à la courbe de codage). *yojE* serait donc bien une CDS complète et fonctionnelle.
B) Le fragment Bs0306.1 '*newAGC*' correspond à la partie C-terminale du gène codant un *multidrug-efflux transporter*.

(NG issus de la ligne BACSU*, « ancienne » réannotation, TAB. 10.2 p. 295) : 39 CDS '*newAGC*' et 9 CDS '*ambiguousAGC*' ont été choisies respectivement parmi 50 CDS '*newAGC*' et 66 CDS '*ambiguousAGC*'. On peut séparer ces résultats en trois catégories.

1. *CDS complètes* : ce sont des CDS qui ne sont pas impliquées dans des décalages du cadre de lecture et dont la longueur est supérieure à 250 pb
2. *Petites CDS complètes* : ce sont des CDS qui ne sont pas impliquées dans des cas de *frameshift* mais dont la longueur est inférieure ou égale à 250 pb.
3. *Fragments de CDS* : ce sont des CDS qui mettent généralement en évidence des mutations (décalages du cadre de lecture, codon de terminaison en phase). Ces mutations peuvent révéler des erreurs de séquençage, des mécanismes de régulation (*frameshift* programmé), des pertes de fonction (pseudogène), des duplications avec perte de fonction, d'autres réarrangements

imparfaits comme une fusion–fission incomplète ou une délétion partielle. Les pseudogènes doivent absolument être annotés car ils signent l'évolution des chromosomes bactériens.

Nous dénombrons ainsi 12 *CDS complètes*[7], 14 *petites CDS complètes*[8] et 22 *fragments de CDS*.

gene	L	Pc	O	Comment	verdict
ybcO	168	0,063	N	Pc=0.629 mat III	keep
ydaQ	291	0,016	Y	in front of N Bs0435.1	remove
ydiN	216	0,001	N	Pc=0,211 mat III in prophage 3 region (FS)	?
ydiQ	189	0,005	Y	in front of N Bs0608.1 (FS) in prophage 3 region	remove
yfmH	72	0,022	N	Pc=0,183 mat III, FS, peptide leader?	?
yfmA	168	0,007	N	in front of unique AMIGene BsuCDS0826499R (FS)	remove
yflD	138	0,024	N	Pc=0,256 mat III, FS, peptide leader?	?
yhdS	111	0,000	N	opposite strand to the previous and the next CDS	remove
yheJ	162	0,041	N	Pc=0,371455 mat II	keep
yhaY	252	0,000	N		remove
yjcB	216	0,008	Y	overlaps A Bs1179.1, FS?	?
yjcE	303	0,001	Y	in front of A Bs1183.1	remove
ynxB	291	0,084	N	Pc=0,418 mat III	keep
ynaG	276	0,014	N	in front of unique AMIGene BsuCDS1884545R (FS)	remove
ypuE	153	0,004	N		remove
yrkS	165	0,019	N	included in repeat region (2701107..2701478)	remove
yrkM	108	0,000	N		remove
yrkG	111	0,048	N	close to *yrkH* (FS)	remove
yrkB	147	0,008	N	Pc=0,174 mat III, FS ?	?
ythC	168	0,088	Y	Pc=0,458 mat II	keep
ywdC	429	0,001	Y	in front of N Bs3800.1	remove
yxjJ	264	0,062	N	Pc=0,493 mat I	keep
yxiL	201	0,100	Y	Pc=0,179 mat I, overlaps A Bs3911.1 and A Bs3912.1, FS?	?

TAB. 10.10 – CDS '*wrongBank*' chez *B. subtilis* (pour la légende voir TAB. 10.9 p. 324)

Parmi les *CDS complètes*, la justesse de prédiction des CDS Bs0435.1 et Bs3800.1 est confirmée par leur recouvrement avec des CDS '*wrongBank*' respectivement *ydaQ* et *ywdC*. Le cas Bs3800.1 ressemble à un « oubli » d'annotation. En effet, il existe deux entrées dans TrEMBL-EBI O31859 *GENE :YOJE OR YOJD* et O68260 *GENE :YOJE* chez *B. subtilis* qui pointent à tort sur le même nom de gène. Pour lever cette confusion, O31859 et O68260 devraient correspondre respectivement aux protéines YojD et YojE. Dans les annotations de référence de la base SubtiList (version R16.1 [Moszer *et al.*, 2002]), la CDS *yojE* devrait être renommée en *yojD* et la « vraie » CDS *yojE* devrait être ajoutée (FIG. 10.8 A p. 325). Bs3126.1 ne présente pas de similitude avec les séquences de la SWALL, ce qui la rend d'autant plus intéressante.

Cinq *petites CDS complètes* ne présentent pas de similitude avec les séquences de la SWALL (statut '*ambiguousAGC*'). La justesse de prédiction de l'une d'entre elles, Bs1183.1, est confirmée

[7]Bs3568.1, Bs2122.1, Bs2922.1, Bs3633.1, Bs3126.1, Bs2359.1, Bs0435.1, Bs2962.1, Bs3800.1, Bs1838.1, Bs1947.1 et Bs3265.1.

[8]Bs3284.1, Bs0889.1, Bs2771.1, Bs2699.1, Bs2923.1, Bs2745.1, Bs1381.1, Bs3194.1, Bs1999.1, Bs3901.1, Bs1183.1, Bs1743.1, Bs4056.1 et Bs1264.1.

par son recouvrement avec la CDS '*wrongBank*' *yjcE*. Neuf *petites CDS complètes* présentent des similitudes dans la SWALL et cinq d'entre elles sont similaires à des *petites CDS complètes* de *B. halodurans*.

La catégorie des *fragments de CDS* est aussi importante car elle permet d'annoter correctement les pseudogènes, voire de mettre en évidence des *frameshift* programmés et de reconstruire les CDS complexes (type *cCDS* composé de *fCDS* dans la base PkGDB ; voir p. 193). Seule une validation expérimentale permet de trancher entre un cas de pseudogène et un cas *frameshift* programmé lorsque la CDS complexe est constituée de deux fragments (à partir de trois, il est raisonnable de la considérer comme un pseudogène). Les réannotations R16.1 ne révèlent que quatre fCDS (un codon de terminaison en phase pour *ywtF* et un décalage du cadre de lecture pour *prfB*). Ces nouveaux fragments permettent de mettre en évidence d'autres points de mutation dans le chromosome. Dans l'exemple de la figure 10.8 B p. 325, il est évident que la CDS *mdr* ne peut être annotée comme une protéine complète et fonctionnelle.

Le tableau 10.10 p. 326 synthétise les résultats de 23 *Gene Not Found* (GNF) '*wrongBank*' choisis parmi 29. On peut séparer ces résultats en trois catégories : *à conserver*, *à confirmer* et *à éliminer*. Les cinq CDS de la catégorie *à conserver* correspondent « actuellement » à des CDS communes aux annotations des banques et aux prédictions *AMIGene* selon le nouveau pipeline (ces CDS sont généralement mieux prédites par les matrices II ou III ; Tab. 10.10 p. 326). Les six CDS *à confirmer* correspondent généralement à des cas possibles de décalage du cadre de lecture mais qui ne peuvent actuellement être confirmés par des résultats de similitude (les CDS de statut '*wrongBank*' sont par définition '*NOSIM*'). Il reste donc 12 CDS *à éliminer*.

10.5 Utilisation de ces résultats au sein du projet *HAMAP*

Une application concrète du processus de réannotation des génomes procaryotes complets est l'intégration de résultats de la présente étude à la banque Swiss-Prot (Table 1 de l'*Article III* p. 289). Plus précisément, un des objectifs de ce travail est de fournir des annotations homogènes au projet d'annotation des protéomes microbiens, manuelle, automatique et de grande qualité (*HAMAP* [Gattiker *et al.*, 2003]), qui consiste à réannoter automatiquement un pourcentage significatif des protéines de TrEMBL-EBI correspondant aux fichiers d'annotation des réplicons procaryotes complets, afin d'alimenter Swiss-Prot [Boeckmann *et al.*, 2003]. Le protocole d'annotation d'*HAMAP* se différencie de nombre de ceux existant actuellement en annotation automatique par le fait qu'il ne tente pas d'inférer quoi que ce soit à partir de similitudes distantes (voir p. 157). Les programmes développés sont spécifiquement ajustés pour faire ressortir les protéines excentriques (*e.g.* paralogues, fragments, fonction biologique inattendue) ; puis une annotation manuelle méticuleuse est réalisée sur ces protéines. Les résultats présentés ici constituent une contribution au projet *HAMAP* puisqu'ils sont issus d'un processus de réannotation semi-automatique qui nécessite une certaine expertise garantissant la qualité des résultats. Comme le montre le tableau des *potential New Gene* intégrés dans Swiss-Prot (Tab. 10.11 p. 328), des efforts sont fournis pour complé-

Species	Acces	Name	Description	L
Aeropyrum pernix	P58077	RPL21E|APE0548.1	50S ribosomal protein L21e.	107
	P58085	RPL29P|APE0362.1	50S ribosomal protein L29P.	66
	P58026	RPL34E|APE0978.1	50S ribosomal protein L34e.	95
	P59472	RPL39E|APE1087.1	50S ribosomal protein L39e.	51
	P58322	APE0277.1	Hypothetical UPF0175 protein APE0277.1.	89
Aquifex aeolicus	P58413	RPME|AQ_873.1	50S ribosomal protein L31.	68
Archaeoglobus fulgidus	O29720	GYRB|AF0530	DNA gyrase subunit B (EC 5.99.1.3).	632
	P58023	AF0072.1	Hypothetical UPF0150 protein AF0072.1.	66
	P58014	AF0739.1	Hypothetical UPF0146 protein AF0739.1.	130
	P58016	AF2370.1	Hypothetical UPF0147 protein AF2370.1.	87
	P58024	AF2407.1	Hypothetical protein AF2407.1.	64
Chlamydia trachomatis	P58001	XSEB|CT329.1	Probable exodeoxyribonuclease VII small subunit (EC 3.1.11.6)|(Exonuclease VII small subunit).	72
Helicobacter pylori	P57798	HP0585.1	Hypothetical protein HP0585.1.	76
Methanococcus jannaschii	P58018	MJ1351.1	Hypothetical protein MJ1351.1.	364
Mycoplasma genitalium	P58061	SECG|MG103.1	Probable protein-export membrane protein secG.	77
	Q49329	MG269.1	Hypothetical protein MG269.1.	50
Pyrococcus horikoshii	P58503	RPS17E|PH1316.1	30S ribosomal protein S17e.	67
	P58189	RPL31E|PH0529.1	50S ribosomal protein L31e.	95
	P58746	RPS24E|PH1909.1	30S ribosomal protein S24e.	99
	P58078	RPS27E|PH1939.1	30S ribosomal protein S27e.	65
	P58193	PH1771.1	Protein translation factor SUI1 homolog.	99
Thermotoga maritima	P58008	TM0562.1	Hypothetical protein TM0562.1.	192
	P58009	TM1158.1	Hypothetical protein TM1158.1.	240
	P58010	TM1467.1	Hypothetical protein TM1467.1.	168
	P58011	TM1791.1	Hypothetical protein TM1791.1.	129
Treponema pallidum	P58007	TP0409.1	Hypothetical UPF0092 protein TP0409.1.	135
Vibrio cholerae	P58093	VCA0360.1	Hypothetical UPF0156 protein VCA0360.1.	80

TAB. 10.11 – Nouvelles CDS dans Swiss-Prot (les noms d'espèces en vert correspondent aux archées)

ter les catalogues des fonctions biologiques essentielles, comme celles impliquées dans le processus de traduction cellulaire (RL21, RL29 et RL34 d'*A. pernix*) ou dans les systèmes d'exportation (SecG de *M. genitalium*). De nouveaux gènes, ayant des similitudes sur l'ensemble de leur séquence avec d'autres protéines hypothétiques de la banque, sont aussi inclus dans les nouvelles entrées de Swiss-Prot (YD5A de *M. jannaschii*, Y56A et YB5A de *Thermotoga maritima*). Dans le cas GyrB d'*Archaeoglobus fulgidus*, nous avons trouvé une CDS courte qui suggère fortement un décalage du cadre de lecture dans le gène *gyrB*. Bien que le nom de ce gène reste identique à celui donné par la banque nucléique (*i.e.* AF0530), l'entrée Swiss-Prot a été corrigée : la séquence protéique a été allongée (632 aa au lieu de 507).

La course aux séquençage, les différents processus d'annotation, les différentes manières de formater les informations nécessitent une réannotation des génomes complets procaryotes. De nombreux projets en cours œuvrent pour faciliter les tâches de (ré)annotation. Par exemple, le consortium *Gene Ontology*, plus centré sur les génomes eucaryotes, a pour objectif de créer un standard de description des gènes afin d'uniformiser le contenu des annotations [Ashburner *et al.*, 2000]. Aussi, des plates-formes d'exploration des génomes sont développées comme *Genostar* [Durand *et al.*, 2003]. Nous pouvons aussi citer les projets *RefSeq* (curation automatique et manuelle des annotations originales de GenBank) et TPA (réannotation expérimentale d'annotations de GenBank ; voir p. 61). Néanmoins, la validation définitive et incontestable des cas difficiles ne peut se faire que par l'expérimentation comme dans le cas du gènes *secE* d'*H. pylori* (*Article IV* p. 316).

Chapitre 11

Annotation et exploration de groupes de CDS de génomes d'entérobactéries

Tout au long de ce manuscrit, nous avons décrit des exemples de gènes excentriques tels que des CDS impliquées dans des décalages du cadre de lecture, des CDS contenant des répétions et/ou des CDS de composition atypique (*putative Alien* (pA) [Karlin, 2001]). Nous avons plus particulièrement décrit des îlots génomiques qui peuvent être impliqués dans la virulence des *Protéobactéries* Gamma pathogènes (*Genomic Island* (GI) [Wren, 2000]). Le sujet des îlots de pathogénie bactériens est à l'intersection de deux thèmes de recherche controversés (*PAthogenicity Island* (PAI) [Hacker & Kaper, 2000]) :

1. le transfert horizontal de gènes (*Horizontal Gene Transfert* (HGT)) et

2. les facteurs de virulence chez les bactéries pathogènes.

Dans l'*état de l'art*, nous avons défini le transfert horizontal de gènes (voir p. 46) comme un mécanisme d'évolution chez les bactéries (Fig. 1.3 p. 54 [Lawrence, 2002]), et la virulence bactérienne (Fig. 1.5 p. 57 [Wassenaar & Gaastra, 2001]).
Dans la partie *méthodologie*, nous avons défini et présenté des îlots génomiques (voir p. 191 ; p. 232 et p. 198). Nous avons aussi discuté l'hypothèse d'une classe de gènes ayant un usage des codons synonymes atypiques *i.e.* AT_3 riche et pouvant être issus de transferts horizontaux (voir p. 226 [Friis *et al.*, 2000]).
Enfin, dans cette partie *prédictions biologiques*, nous allons présenter l'annotation du génome d'une entérobactérie entomopathogène, *P. luminescens*, et explorer ses îlots génomiques à la lumière des groupes de synténies partagés ou non avec d'autres entérobactéries (*e.g.* le modèle commensal *E. coli* K-12, l'agent de la peste *Y. pestis* CO92).

11.1 *Article V* : « The genome sequence of the entomopathogenic bacterium *P. luminescens* »

E. Duchaud, C. Rusniok, L. Frangeul, C. Buchrieser, A. Givaudan, S. Taourit, S. Bocs, C. Boursaux-Eude, M. Chandler, J. F. Charles, E. Dassa, R. Derose, S. Derzelle, G. Freyssinet, S. Gaudriault, C. Médigue, A. Lanois, K. Powell, P. Siguier, R. Vincent, V. Wingate, M. Zouine, P. Glaser, N. Boemare, A. Danchin, F. Kunst, F. (2003) Nature biotechnology, 1087-0156.

P. luminescens est une entérobactérie insecticide, antibiotique, antifongique et bioluminescente, qui vit en symbiose dans le tube digestif d'un nématode (voir p. 31). Le génome de *P. luminescens*, un chromosome circulaire contenant un total de 4839 gènes codant des protéines, a été séquencé et annoté par E. Duchaud sous la direction de F. Kunst et P. Glaser, responsables du Laboratoire de Génomique des Microorganismes Pathogènes de l'Institut Pasteur. L'analyse du chromosome nouvellement séquencé de cette entérobactérie entomopathogène, est réalisée en collaboration avec l'INRA-Université de Montpellier II, d'autres équipes du CNRS et de l'Institut Pasteur, et la société Bayer CropScience. Nous participons à cette analyse sur plusieurs aspects :

- Avant la publication du génome, E. Duchaud nous a envoyé son jeu de CDS annotées, que nous avons réannoté avec trois matrices de transitions spécifiques de l'usage des codons synonymes. L'analyse de nos résultats lui a permis de modifier son jeu d'annotation.
- Finalement, quatre classes de gènes (au lieu de trois) ont été définies en fonction de l'usage des codons synonymes (Fig. 7.4 p. 218). Elles ont servi à construire quatre matrices de transitions (Fig. 7.8 p. 232).
- La réannotation avec quatre matrices de transition du jeu de CDS de *P. luminescens* publié, ne produit aucun GNF de statut '*wrongBank*' mais produit une proportion significative de nouveaux gènes potentiellement intéressants (Tab. 10.2 p. 295).
- Enfin, *P. luminescens* fait partie des entérobactéries de la base *Entero*DB qui nous permet de réannoter ces génomes afin d'explorer notamment leur îlots génomiques à la lumière des groupes de synténie (voir p. 195).

L'analyse du génome de *P. luminescens* révèle l'existence de toute une variété de gènes codant des toxines susceptibles de tuer de nombreux insectes. Aucun génome bactérien aujourd'hui séquencé n'avait permis de trouver autant de gènes de toxines entomopathogènes. La toxicité de ces protéines a été vérifiée expérimentalement [ffrench Constant & Bowen, 2000] ; certaines se sont avérées mortelles pour les moustiques (lutte contre les insectes nuisibles pour la santé). De plus, des gènes impliqués dans la symbiose entre cette bactérie et le nématode qui l'abrite ont aussi été découverts. Ce *couple* pourrait être utilisé pour la lutte biologique contre les insectes nuisibles pour l'agriculture.

Enfin, l'analyse de ce génome permet d'identifier toute une gamme de gènes codant des protéines de biosynthèse d'antibiotiques et d'antifongiques, sources potentielles de retombées pour le traitement des maladies infectieuses.

11.2 Ilots génomiques

Les génomes complets de bactéries pathogènes sont très utiles pour comprendre les mécanismes fondamentaux utilisés par les microbes pour causer l'infection et la maladie [Finlay & Falkow, 1997, Hood, 1999]. Les facteurs de virulence des bactéries pathogènes (adhésines, toxines, invasines, système de sécrétion de protéines, système de capture du fer, etc.) peuvent être encodés par des régions particulières des génomes appelés îlots de pathogénie [Hacker & Kaper, 2000]. Mais avant de nous intéresser aux îlots de pathogénie, nous allons d'abord nous concentrer sur les îlots génomiques, concept déjà suffisamment complexe.

11.2.1 Approche extrinsèque

Le pionnier de la nouvelle théorie de l'évolution, à la fois verticale et horizontale, est M. Syvanen qui annonce en 1985 que l'universalité du code génétique est une condition nécessaire au mécanisme du transfert horizontal de gènes [Syvanen, 1985]. Il prône l'approche extrinsèque où l'on compare un gène par rapport aux gènes d'autres génomes à la recherche de résultats de similitude inattendus qui sont confirmés dans un contexte évolutif (*e.g.* test de congruence phylogénétique, distribution en gènes entre espèces relativement éloignées, contenu en gènes entre espèces relativement proches ; [Syvanen, 1994]). Le test de congruence entre différentes topologies d'un arbre a été défini pour discuter la possibilité de transfert entre procaryotes et eucaryotes [Smith *et al.*, 1992].

En 1999, N. J. Saunder *et coll.* utilise le fait que le transfert d'ADN entre deux espèces bactériennes proches (*e.g.* phylum des *Protéobactéries*) est guidé par des séquences de capture, espèce-spécifique (*DNA uptake sequence* chez les espèces à gram négatif), pour identifier des gènes qui ont été transférés d'*Haemophilus* aux *Neisseria* [Saunders *et al.*, 1999]. Par exemple, la superoxide dismutase encodée par *sodC* est présente chez *H. influenzae* et *N. meningitidis*, mais pas chez *N. gonorrhoeae*. De plus, les résultats de similitude suggèrent une origine commune aux gènes *sodC* d'*H. influenzae* et de *N. meningitidis*. Enfin, la présence de séquences de capture d'ADN spécifiques de *H. influenzae* chez *N. meningitidis* indique que le transfert horizontal de *sodC* s'est effectué de *H. influenzae* vers *N. meningitidis*. C'est un exemple clair de transfert d'ADN entre espèces dans l'évolution des pathogènes.

En 2000, W. Martin décrit les chromosomes bactériens comme des mosaïques qui ouvrent le challenge de reconstruction d'un arbre des génomes complets procaryotes [Martin, 1999]. Fitch, quant à lui, définit la relation de xénologie entre deux gènes comme une relation d'homologie entre ces deux gènes dont l'ancêtre commun implique un mécanisme de transfert horizontal [Fitch, 2000].

11.2.2 Approche intrinsèque

C'est au début des années 1990 que C. Médigue a initié l'étude du transfert horizontal de gènes chez *E. coli* K-12 [Médigue *et al.*, 1991]. Un échantillon des gènes d'*E. coli* K-12 a été classé en fonction des résultats de l'AFC suivant l'usage des codons synonymes. Une partition en trois classes

a ainsi été obtenue, contenant respectivement 64,2% des gènes en classes I (expression faible ou moyenne), 24,4% en classe II (expression haute et constitutive) et 11,4% en classe III (probablement issus de transfert horizontaux), sur un total de 782 gènes.
Cette méthode relève d'une approche intrinsèque ou paramétrique qui utilise la statistique descriptive pour comparer un (ou un groupe de) gène(s) par rapport aux autres (ou à un autre groupe de) gènes du génome à la recherche de séquences atypiques (*e.g.* composition en bases, fréquence en dinucléotides, usage des codons synonymes).

Par la suite, J. G. Lawrence et H. Ochman ont étudié le transfert horizontal de gènes chez *E. coli* K-12 en combinant cinq critères (approche mixte intrinsèque–extrinsèque [Lawrence & Ochman, 1998]) :

1. Une CDS est initialement qualifiée d'atypique si la déviation de son contenu en G+C1 et en G+C3 est de plus de deux fois la déviation standard associée aux moyennes respectives calculées sur l'ensemble des 4288 gènes d'*E. coli* K-12.
2. Ces CDS atypiques sont alors séparées en deux groupes : celles qui ont un usage des codons synonymes atypique parce qu'elles sont constitutivement et fortement exprimées (*predicted highly expressed* (PHX)) et celles qui ont un usage des codons synonymes atypique parce qu'elles sont issues de transferts horizontaux. Pour cela, on commence par calculer pour chaque gène :
   (a) son biais d'usage des codons synonymes par rapport à l'usage des codons attendus connaissant la composition en bases à la première et la troisième position des codons (un χ^2 sur l'usage des codons),
   (b) son indice d'adaptation des codons (CAI) en référence à un jeu de gènes d'*E. coli* K-12 connus pour être fortement exprimés constitutivement.

   Le χ^2 sur l'usage des codons permet de révéler un usage des codons atypique, même pour un gène qui proviendrait d'un génome de même composition en bases, mais dont l'usage des codons synonymes serait différent. Puis l'on représente le biais d'usage des codons synonymes des gènes en fonction de leur CAI. Les gènes transférés horizontalement sont les gènes atypiques qui présentent à la fois un fort biais d'usage des codons synonymes et un faible CAI.
3. Ces gènes HGT putatifs sont regroupés en fonction de leur colocalisation sur le chromosome afin d'estimer le nombre d'îlots génomiques (*i.e.* le nombre d'événements de transfert).
4. Cette liste de gènes HGT putatifs est examinée pour identifier des gènes natifs connus qui présentent une composition en bases atypique mais aussi pour d'autres raisons comme le contenu en acides aminés du polypeptide encodé. Par exemple, certaines protéines ribosomiques sont riches en lysine, ce qui contribue au faible contenu en G+C de leur CDS.
5. Enfin, une recherche Blast est effectuée sur toutes les CDS pour détecter des similitudes avec des gènes d'espèces proches.

C'est ainsi que J. G. Lawrence et H. Ochman ont estimé à environ 18%, le pourcentage de gènes HGT putatifs chez *E. coli* K-12. Puis en 2000, ils annoncent qu'une partie significative de la diversité microbienne est issue de transferts horizontaux de gènes entre des espèces procaryotes, mêmes

lointaines (*e.g.* entre deux domaines comme les Archaea et les Bactéries). Ces transferts horizontaux ont effectivement changé les caractères écologiques et pathogènes des espèces procaryotes [Ochman *et al.*, 2000]. Les deux auteurs estiment alors à 12,8%[1] la proportion de gènes HGT putatifs chez *E. coli* K-12. Ils précisent que ce nombre est sous-estimé puisque les méthodes fondées sur le contenu en G+C et le biais d'usage des codons synonymes ne peuvent mettre en évidence (i) ni les événements de transfert très anciens tels que la dissémination des synthétases d'ARNt, (ii) ni les transferts entre bactéries dont les génomes ont un contenu en G+C et un biais de codons similaires. Les nouvelles fonctions introduites grâce aux transfert horizontaux concernent par exemple la résistance aux antibiotiques, la virulence, le métabolisme.

En 1998, S. Karlin *et coll.* ont mis au point un nouvel indice : le biais d'usage des codons synonymes entre deux groupes de gènes (*bias in codons* (BC) [Karlin *et al.*, 1998]). Un gène est qualifié de *putative Alien* (pA) s'il remplit deux conditions : (i) si son usage des codons diffère suffisamment de l'usage moyen des codons de tous les gènes (*e.g.* $BC(g \mid all) \geq 0,52$) et (ii) si son usage des codons diffère suffisamment de l'usage moyen des codons des gènes codant des protéines ribosomiques (*e.g.* $BC(g \mid RP) \geq 0,45$). Cette dernière condition permet d'écarter les gènes qui ont un usage des codons synonymes atypique parce qu'ils sont fortement exprimés de façon constitutive. Par cette méthode, S. Karlin *et coll.* détectent seulement 3% de gènes HGT putatifs chez *E. coli* K-12.
En 2001, cet auteur utilise cet indice dans une méthode de prédiction d'îlots génomiques (ou groupe de gènes atypiques *Genomic Island* (GI)) et d'îlot de pathogénie chez diverses génomes. Cette méthode utilise une fenêtre glissant le long du chromosome pour mesurer cinq critères : (i) le pourcentage en G+C (ii) le profil de signature génomique fondé sur les biais en dinucléotides, (iii) les biais d'usage des codons synonymes par rapport à l'usage moyen de tous les gènes (*e.g.* $BC(g \mid all) \geq 0,52$), (iv) les biais d'usage en acides aminés et (v) les biais d'usage des codons synonymes par rapport à l'usage moyen de différentes classes de gènes (*e.g.* $BC(g \mid RP) \geq 0,45$; [Karlin, 2001]).

11.2.3 Latéralistes et verticalistes

C'est en 2001 que la polémique sur le transfert horizontal de gènes éclate. S. Guindon et G. Perrière reprennent les études sur *E. coli* K-12 de J. G. Lawrence et H. Ochman [Lawrence & Ochman, 1998] et de S. Karlin *et coll.* [Karlin *et al.*, 1998]. Ils définissent quatre jeux de référence (CDS d'*E. coli* K-12 de longueur supérieure à 150 pb [Guindon & Perrière, 2001]) : (i) un jeu de 317 gènes HGT putatifs, (ii) un jeu de 17 gènes hautement exprimés, (iii) un jeu de gènes standards (4254 gènes du fichier des annotations du chromosome complet d'*E. coli* K-12) et (i) un jeu de 3937 (4254 - 317) gènes non HGT. Pour le groupe de gènes HGT putatifs et pour le groupe de gènes non HGT, S. Guindon et G. Perrière mesurent cinq indices : (i) l'indice d'adaptation des codons du groupe par rapport au jeu de gènes standard ($CAI(g \mid all)$), (ii) l'indice d'adaptation des

[1] La diminution de 18 à 12,8% peut s'expliquer s'ils ont employé la même méthode qu'en 1998 mais sans l'étape 5 de détection de HGT par une approche extrinsèque.

codons du groupe par rapport au jeu de gènes hautement exprimés ($CAI(g \mid PHX)$), (iii) le biais de codons du groupe par rapport au jeu de gènes standard ($BC(g \mid all)$), (iv) le biais de codons du groupe par rapport au jeu de gènes PHX ($CAI(g \mid PHX)$) et (v) le contraste en G+C aux trois positions de codons ($G+C3c$). Les valeurs de ces cinq mesures entre les deux groupes HGT – non HGT sont statistiquement différentes. Parmi ces cinq indices, il apparaît que c'est le $G+C3c$ qui est la meilleure mesure pour séparer les gènes HGT des gènes non HGT. De plus, de manière surprenante, ces deux auteurs s'aperçoivent que les gènes HGT putatifs ne sont pas symétriquement atypiques : il y a plus de gènes HGT putatifs qui sont G+C3 pauvres, que de gènes HGT putatifs G+C3 riches. Ils se demandent si les études précédentes n'ont pas surestimé le nombre de gènes HGT putatifs, en qualifiant d'HGT des gènes qui ont un usage de codons atypiques non pas parce qu'ils ont été transférés mais parce qu'ils se situent préférentiellement autour du terminus de la réplication (A+T riche [Deschavanne & Filipski, 1995]).
Depuis, G. Perrière a révisé son jugement : en effet, V. Daubin et G. Perrière annoncent finalement que les putatifs HGT récents sont A+T riches (indépendamment de la composition moyenne en GC du génome donneur) et que peu d'entre eux sont ratés par les analyses compositionnelles [Daubin *et al.*, 2003a].

L. B. Koski *et coll.* reconnaissent que le transfert horizontal est un mécanisme important de l'évolution, mais ils pensent que la composition en bases et le biais de codons sont de pauvres indicateurs du transfert horizontal [Koski *et al.*, 2001]. En partant des mêmes relations de correspondance et de proximité que pour la prédiction de groupes de synténie, ils définissent 2728 orthologues putatifs dont la position relative est conservée entre les génomes d'*E. coli* K-12 et de *S. enterica* serovar Typhi CT18, et un groupe de 1144 gènes d'*E. coli* K-12 qui divergent de manière inattendue de leur contre-partie chez *S. enterica* serovar Typhi CT18. Ils pensent que ce groupe de gènes contient les gènes HGT putatifs et des orthologues qui ont été réarrangés ou délétés. Ils trouvent de nombreux gènes de composition en bases normale qui n'ont apparemment pas d'orthologue chez *S. enterica* serovar Typhi CT18 et de nombreux gènes de composition atypique qui ont bien des orthologues positionnels. Ainsi les gènes de classe III ne sont pas tous issus de transferts horizontaux et les gènes de classe I ne sont pas tous natifs.

M. Ragan reconnaît lui aussi que le transfert horizontal de gènes entre en compétition avec la transmission verticale dans l'évolution des microbes, mais il révèle qu'aucune des approches mises en œuvre jusqu'à présent pour les détecter ne donne de résultats cohérents [Ragan, 2001a]. L'approche phylogénétique (incongruence entre les arbres) est en principe la plus rigoureuse pour identifier d'anciens HGT mais les méthodes de reconstruction d'arbres sont soumises à de nombreux artefacts [Syvanen, 1994]. L'approche statistique de comparaison d'un gène avec les autres gènes du génome (composition atypique) est fondée sur l'hypothèse que les gènes de composition atypiques A+T riches sont issus de transferts horizontaux mais rien ne permet de le démontrer. Y aurait-il une autre raison que le transfert horizontal qui pourrait expliquer qu'une composition atypique de gènes puisse surgir et être maintenue dans un génome ? L'approche statistique de comparaison d'un gène avec des gènes similaires d'autres génomes (distribution des gènes) souffre du choix

dans les seuils pour définir si deux séquences sont similaires ou non. C'est pourquoi il propose une nouvelle méthode qui appartient à l'approche qui étudie la distribution des gènes dans les génomes [Ragan & Charlebois, 2002]. Il prend soin de définir deux seuils de similitudes : une E-value inférieure à *e.g.* 10^{-20} permet d'établir que deux séquences sont similaires et une absence de E-value inférieure à 10^{-5} permet d'établir que la séquence ne présente pas de similitude. De plus, la méthode de M. Ragan repose sur l'existence d'un arbre phylogénétique dont il ne connaît pas la topologie. Ainsi il peut comparer la distribution des gènes entre les phyla et entre les domaines. Les gènes qui sont uniformément distribués entre les phyla et les domaines sont probablement transmis verticalement. Les gènes d'un phylum qui ne s'alignent qu'avec exactement un autre phylum sont généralement absents des autres domaines : ce sont des gènes HGT putatifs.

Enfin, les plus verticalistes des chercheurs, C. G. Kurland *et coll.*, pensent que nous exagérons la part du transfert horizontal dans l'évolution des microbes [Kurland *et al.*, 2003]. Plus précisément, ils pensent que le HGT a influencé significativement l'évolution des organismes primitifs [Woese, 2000]. Chez les organismes modernes, ils suggèrent que l'étendue et la fréquence des HGT sont plus contraintes par les barrières sélectives. En conséquence, la plupart des événements HGT influencent peu la phylogénie des génomes. Bien que le HGT ait d'importantes conséquences sur l'évolution, les lignées darwiniennes classiques semblent être le mode dominant de l'évolution des organismes modernes.

Ainsi, les chercheurs sont globalement d'accord pour dire que le transfert horizontal de gènes joue ou a joué une part importante dans l'évolution des microbes ; mais il n'existe pas de données consensuelles [Ragan, 2001b]. Dans le détail, les opinions divergent, allant de l'idée que le HGT a un impact mineur sur l'évolution et la diversité des microbes, à celle que le HGT est tellement fréquent qu'il entrave toute tentative d'élucider l'évolution des microbes (plus précisément la phylogénie des organismes à partir des comparaisons de séquences [Lawrence & Hendrickson, 2003]). A court terme, il semble raisonnable de qualifier de gènes HGT putatifs l'ensemble des gènes prédits HGT par différentes méthodes [Lawrence & Ochman, 2002]. Ainsi, en réconciliant les résultats d'une approche extrinsèque et d'une approche intrinsèque, ils estiment à 24,5% la proportion de gènes HGT putatifs chez *E. coli* K-12. Dans sa réponse, M. A. Ragan souligne que ce nombre est sous-estimé puisqu'il ne concerne que les HGT *récents*. A long terme, il devient donc urgent de développer de nouvelles méthodes (modèles de simulation [Gogarten *et al.*, 2002], classifications fondées sur les réseaux et non sur les arbres) et surtout de valider expérimentalement des transfert horizontaux de différents gènes entre différents génomes ; ceci à afin de pouvoir répondre à quatre questions fondamentales et de trouver un consensus [Lawrence & Hendrickson, 2003] :

1. Comment le HGT influence l'histoire évolutive de différents gènes (*e.g.* les gènes codant la résistance aux antibiotiques semblent plus fréquemment transférés que les gènes codant les protéines ribosomiques [Woese, 2000]) ?
2. Comment l'impact du HGT diffère entre les différentes lignées (*e.g.* certaines bactéries comme *H. influenzae* et *N. meningitidis* sont naturellement transformables relativement à d'autres

comme *E. coli* K-12 [Saunders *et al.*, 1999]) ?

3. Comment obtenir des conclusions robustes quant à la présence ou à l'absence de HGT ?
4. Comment intégrer le HGT dans le continuum des exchanges génétiques pour que cela mène à des concepts de microbiologie qui ont un sens ? Autrement dit, à quelle distance phylogénétique passe-t-on de la recombinaison homologue entre espèces proches à la recombinaison illégitime entre espèce éloignées (HGT) ? et à quelle distance phylogénétique, la recombinaison illégitime entre espèces éloignées n'est-elle plus possible ?

11.2.4 Outils sur le Web

C'est au début de ce troisième millénaire que sont nés les premiers outils en ligne d'analyse d'îlots génomiques et de pathogénie.

Méthodes

Parmi les programmes d'analyse d'îlots génomiques et d'îlots de pathogénie disponible sur le Web, nous trouvons par exemple :

- Le serveur Web *GenomeAtlases* qui permet de visualiser les répétitions, les caractéristiques structurales de l'ADN (*e.g.* courbure, flexibilité de la double hélice d'ADN, énergie d'empilement) et la composition en base des réplicons [Friis *et al.*, 2000, Pedersen *et al.*, 2000].
- Le serveur *Enterix* qui permet de visualiser, à différentes échelles, les paires de génomes alignés, entre un génome pivot d'entérobactérie et chacun des génomes choisis par l'utilisateur dans la même classe des *Protéobactéries* Gamma (Fig. 1.2 p. 53 [Florea *et al.*, 2000, Florea *et al.*, 2003]).
- Le serveur *IslandPath* qui aide à la détection d'îlots génomiques chez les procaryotes [Hsiao *et al.*, 2003]. La méthode combine quatre critères : (i) le pourcentage en G+C de chaque gène (pas de fenêtre glissante), (ii) le biais en dinucléotides de chaque groupe de six gènes, (iii) la présence de gènes de mobilité (*e.g.* transposases, intégrases) et (iv) la présence de gènes d'ARNt.

Ces méthodes permettent de parcourir rapidement un génome à la recherche de régions excentriques (Fig. 7.1 p. 198 et Fig. 10.2 p. 296), potentiellement intéressantes, dont le rôle et l'origine peuvent être caractérisés ultérieurement par des analyses bioinformatiques et/ou par des validations expérimentales.

Ressources

Il existe peu de bases de données publiques sur la pathogénie bactérienne (la compagnie biopharmaceutique Genome Therapeutics a développée une base commerciale multigénome PathoGenome[2]). Nous pouvons tout de même citer la banque d'HGT putatifs des génomes procaryotes

[2] `http://www.genomecorp.com/programs/pathogenome.shtml`

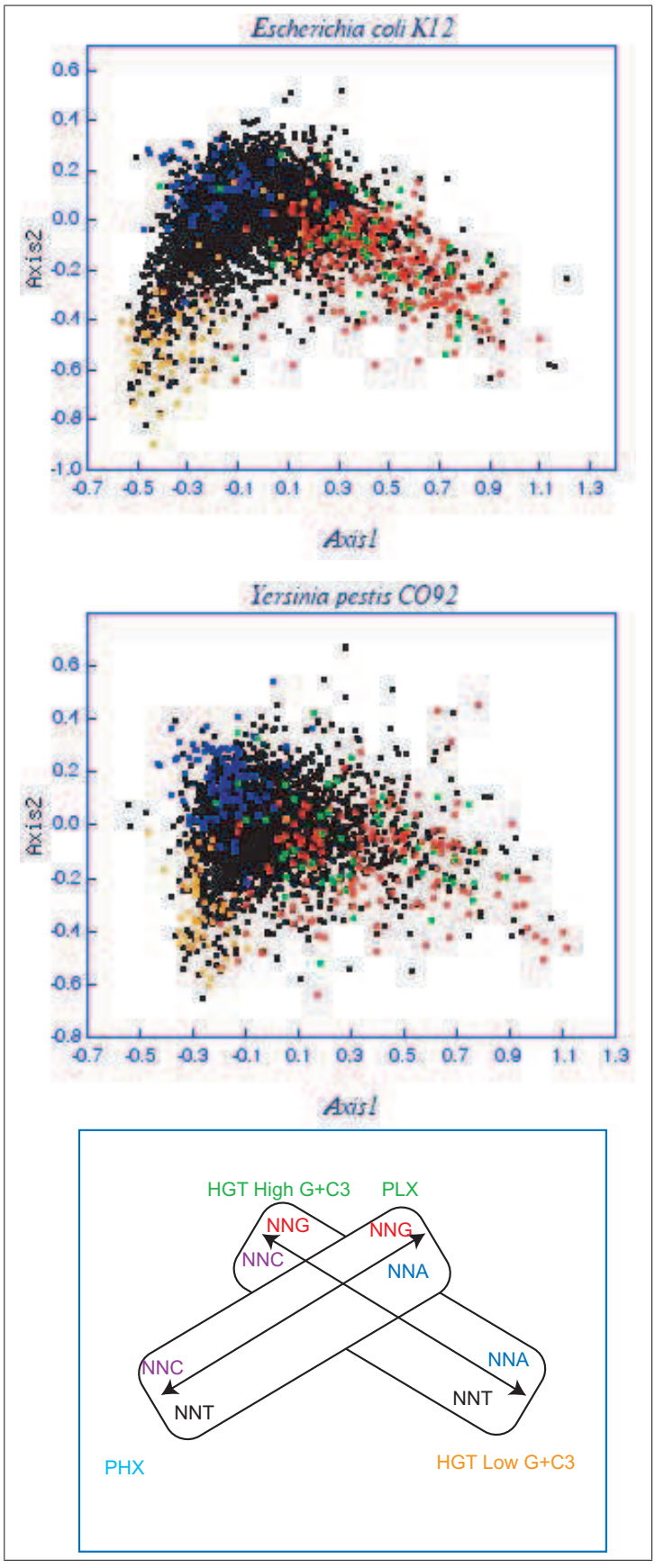


FIG. 11.1 – Gènes issus de transferts horizontaux

Ces représentations de CDS annotées ($L > 300$pb) en fonction de leurs coordonnées sur les deux premiers axes de l'analyse factorielle des correspondances de l'usage relatif des codons synonymes ont été téléchargées à partir du site d'HGT-DB [Garcia-Vallve *et al.*, 2003]. Red, blue and green squares correspond to genes proposed as being acquired by HGT and included in regions with a low, high and non-deviated G+C content, respectively. Yellow cross diagonals (x) correspond to putative highly expressed genes.

complets [Garcia-Vallve *et al.*, 2003]. Les prédictions sont fondées sur une combinaison de plusieurs critères calculés pour chaque gène : (i) le G+C et le G+C3, (ii) l'usage des codons et l'usage relatif des codons synonymes (RSCU), (iii) le contenu en acides aminés. La moyenne et la déviation standard de ces paramètres sont aussi calculée sur l'ensemble des gènes de longueur supérieure à 300 pb. Brièvement, un gène est qualifié d'atypique dans son contenu en G+C ou dans son usage des codons, si la valeur caractérisant le gène dévie de plus de 1,5 fois la déviation standard associée à la moyenne calculée sur l'ensemble des gènes du génome. Pour qu'un gène soit prédit comme un gène HGT putatif (Fig. 11.1 p. 339), il faut : (i) que sa longueur soit supérieure à 300 pb, (ii) que son usage des codons et son contenu en G+C soient atypiques et (iii) que son contenu en acides aminés ne dévie pas de la composition moyenne des produits des gènes du génome.
Islander est une base multigénome contenant des îlots génomiques prédits à partir de séquences chromosomiques bactériennes complètes [Mantri & Williams, 2004]. Un îlot code souvent un gène d'intégrase qui est responsable de l'intégration de l'îlot en un site spécifique du génome. Généralement, les sites spécifiés par les intégrases sont des gènes d'ARNt et l'îlot coupe le gène d'ARNt quand il s'intègre. Cependant, l'îlot porte aussi une séquence qui remplace la portion manquante, restaurant un gène d'ARNt intact. Donc un îlot est souvent marqué par un gène d'ARNt à une extrémité, et un fragment de ce gène à l'autre extrémité. Les îlots de la base Islander sont prédits en utilisant ce principe à travers la procédure suivante :

1. Recherche de gènes d'ARNt et d'ARNtm sur le génome bactérien complet.
2. Recherche d'intégrases.
3. Chaque gène prédit à la première étape est utilisé comme « requête » lors d'une recherche de similitude Blast sur le génome. Le gène « requête » et le fragment « sujet » délimitent les bornes d'un îlot candidat.
4. Chaque îlot candidat doit passer par une série de filtres. Les îlots qui répondent aux critères suivant sont éliminés :
   - L'îlot candidat ne contient pas d'intégrase.
   - Le fragment « sujet » délimitant l'îlot recouvre une CDS annotée de longueur supérieure à 300 pb.
   - Le fragment « sujet » correspond à un ARNt annoté.
   - La longueur de l'îlot candidat est supérieure à 200 kb.
   - Le fragment « sujet » ne s'étend pas jusqu'à une des extrémités du gène « requête ».
   - La configuration du fragment « sujet » par rapport au gène « requête » n'est pas cohérente[3].
   - Le gène « requête » et le fragment « sujet » sont en sens contraire.
5. Les îlots candidats restants qui partagent une même intégrase sont fusionnés et chargés dans la base Islander.

[3]Par exemple, en 3′ du gène « requête », on trouve un fragment « sujet » correspondant à l'extrémité 5′ du gène « requête » (au lieu de trouver un fragment correspondant à l'extrémité 3′ du gène).

11.2.5 Exemples d'îlots génomiques

Nous avons développé et testé une méthode d'annotation relationnelle de prédiction d'îlots génomiques (*e.g.* îlots de pathogénie, îlots de résistance aux antibiotiques,'îlots métaboliques) à partir des annotations syntaxiques et fonctionnelles d'un chromosome complet (résultats non présentés [Bocs *et al.*, 2001]). Il s'agit dans un premier temps de rassembler un maximum de données pertinentes pour la détection d'îlots génomiques : les ARNt, les gènes de mobilité (intégrase de phage, transposase d'IS et de transposon), les caractéristiques des CDS (*e.g.* G+C3, classe d'usage des codons synonymes), les décalages du cadre de lecture, les régions répétées, les annotations fonctionnelles des polypeptides encodés par les CDS (*e.g.* recherche de mots-clés dans le produit). Ces informations sont stockées dans des tables de PkGDB comme [Genomic_Object], [tRNA_Scan], [ProFed], [Codon_Wcluster], [Nosferatu].
Il suffit alors, dans un second temps, d'interroger la base afin de combiner habilement ces informations pour reconstruire des groupes de gènes excentriques colocalisés sur le chrosomome, susceptibles d'avoir été acquis par transferts horizontaux et dont les produits peuvent être impliqués dans des fonctions importantes pour les propriétés d'adaptation de la bactérie à de nouveaux milieux. Les îlots sont stockés dans la table [Genomic_Island] de PkGDB.

Nous avons aussi défini des jeux de référence d'îlots génomiques à partir des annotations des banques. Ainsi, les îlots génomiques annotés (et non prédits) d'*E. coli* K-12 [Serres *et al.*, 2004], de *P. luminescens* (*Article V* p. 332) et de *Y. pestis* CO92 [Parkhill *et al.*, 2001b] ont été stockés dans la table [Genomic_Island] de PkGDB.

Les îlots génomiques annotés ou prédits de [Genomic_Island] peuvent être visualisés de différentes façons.

Vue d'ensemble

Une première façon de visualiser les îlots génomiques consiste à construire une carte circulaire du chromosome complet, par exemple en utilisant les données de PkGDB dans le logiciel GenVision (Fig. 11.2 p. 343). En partant de la périphérie et en allant vers le centre, les cercles représentent différents objets génomiques annotés :

1. les CDS ($L > 200$pb) coloriées en fonction de leur classe d'usage des codons synonymes (I vert, II cyan, III orange et IV magenta),
2. les *frameshifts* (en orange), les IS (en violet) et les ARNt (en vert foncé),
3. les îlots génomiques (en saumon),
4. le GC_3 le long du chromosome (fenêtre de 1000 pb et pas de 200 pb), en bleu si la valeur est inférieure à 1,5 fois la déviation standard calculée par rapport à l'ensemble des valeurs du chromosome, en rouge si la valeur est supérieure à 1,5 fois la déviation standard et en jaune, le reste du temps.

On remarque qu'il existe des îlots génomiques (en rose saumon sur le cercle3) chez ces trois génomes, que l'entérobactérie soit pathogène ou non. Le génome de *P. luminescens* est celui qui semble

contenir le plus de gènes HGT putatifs, ce qui corrobore le fait que ce soit le génome séquencé qui possède à ce jour le plus de gènes dont les produits interviennent dans la biosynthèse de toxines entomopathogènes, d'antibiotiques et d'antifongiques. Bien entendu, nombre de gènes HGT putatifs codent des polypeptides de fonction encore totalement inconnue (ORFan).

De plus, les îlots peuvent contenir des régions riches en G+C (en rouge) et/ou pauvres en G+C (en bleu sur le cercle4), voire des régions sans biais de G+C (relativement au contenu moyen du génome ; en jaune). De même, ils peuvent contenir des gènes des trois ou quatre classes suivant l'usage des codons synonymes du génome (Fig. 7.4 p. 218). Par exemple, chez *E. coli* K-12 et *Y. pestis* CO92, trois classes de gènes ont été définies : la classe I des gènes typiques (en vert sur le cercle1), la classe II des gènes prédits hautement exprimés (en cyan) et la classe III des gènes HGT putatifs (en orange). Chez *P. luminescens*, quatre classes ont été définies : la classe I des gènes typiques du brin précoce, la classe II des gènes prédits hautement exprimés, la classe III des gènes HGT putatifs et la classe IV des gènes typiques du brin tardif (en magenta). Ainsi, nous avons vu que les îlots d'*E. coli* K-12 se répartissaient préférentiellement en classe I et III alors que ceux de *P. luminescens* semblent se répartir uniformément dans les quatre classes (voir p. 226).

Ces génomes contiennent de nombreux ARNt (en vert foncé sur le cercle2), IS (en violet) et décalages du cadre de lecture (en orange). Il semble que le génome de *Y. pestis* CO92 soit celui qui contienne le plus d'IS et de décalages du cadre de lecture, puis vient *P. luminescens* et enfin *E. coli* K-12. Les îlots de *Y. pestis* CO92 et de *P. luminescens* que nous allons inspecter plus en détail sont indiqués par un quartier gris et une flèche marron sur la figure 11.2 p. 343.

Vue plongeante

L'interface cartographique *MaGe*, développée au dessus de PkGDB, nous permet de visualiser ces îlots à la lumière des synténies. L'îlot choisi pour *Y. pestis* CO92 (en jaune sur la figure 11.3 A p. 346) est annoté comme vestige de phage [Parkhill *et al.*, 2001b] et a été présenté lors de notre communication sur des stratégies d'investigation de la virulence de *Y. pestis* CO92 utilisant la génomique comparative (plus précisément la prédiction de synténie [Labarre *et al.*, 2003, Chain *et al.*, 2004]). La recherche de régions uniques à partir de la prédiction de groupe de synténie nous permet de conclure que cet îlot est absent chez *Y. pseudotuberculosis*, *E. coli* O157:H7 EDL933 et *E. coli* K-12, et partiellement présent chez *P. luminescens*. Il est intéressant puisqu'il est spécifique de *Y. pestis* CO92 par rapport à *Y. pseudotuberculosis* et pourrait donc être un îlot de pathogénie à l'origine de la haute virulence[4] de *Y. pestis* CO92 relativement à *Y. pseudotuberculosis*. Il possède des caractéristiques classiques d'un îlot génomique :

- Il est borné en 5′ par un gène d'ARNt et en 3′ par un gène de petit ARN stable.
- Il contient à la fois des gènes de transposase d'IS (*e.g.* YPO1085, YPO1086, YPO1099) et des gènes d'intégrase de phage (*e.g.* YPO1086a, YPO1098).

[4]Personne ne sait si la haute virulence de *Y. pestis* CO92, relativement à *Y. pseudotuberculosis* est due à une perte de fonctions ou à un gain.

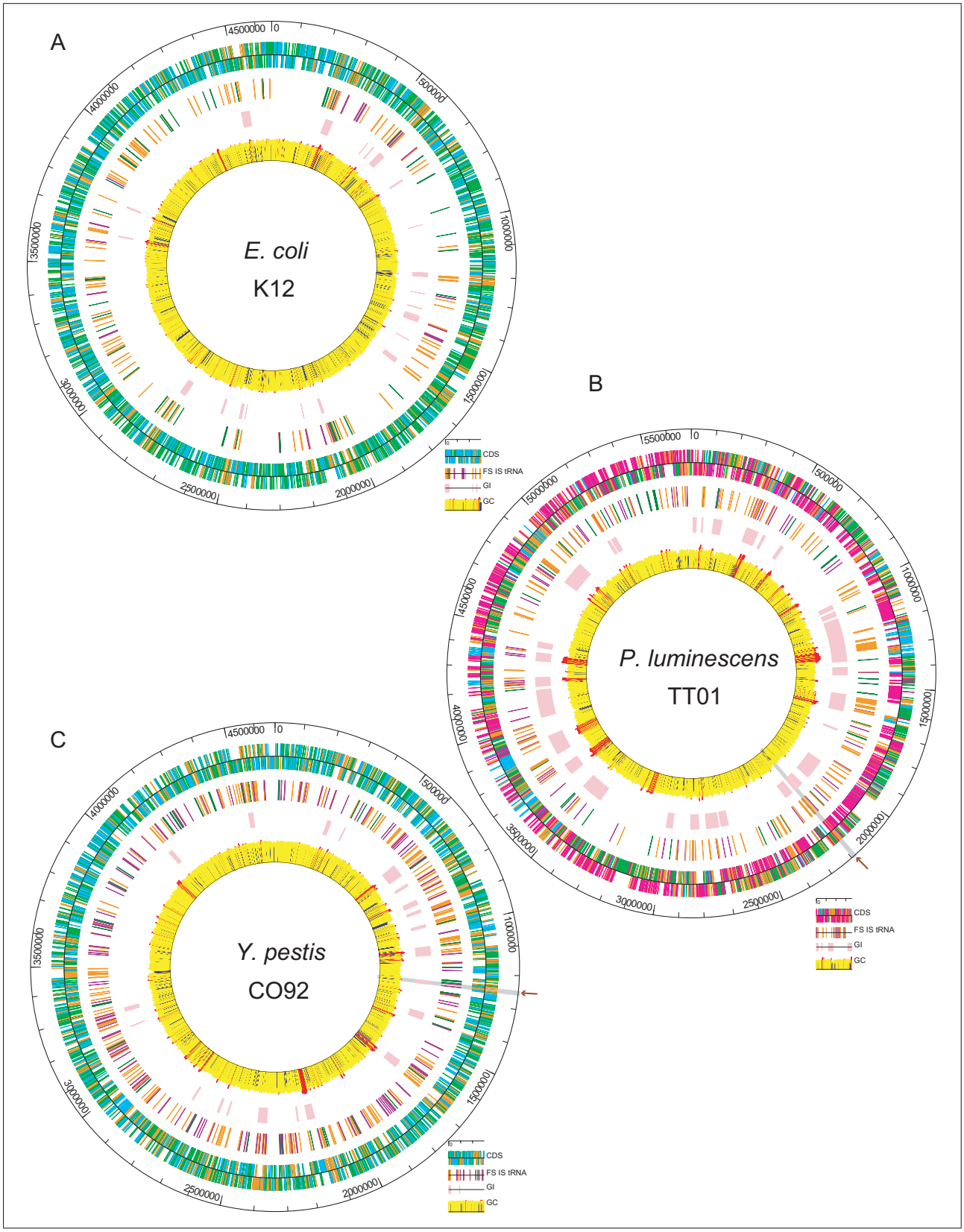


FIG. 11.2 – Visualisation d'îlots génomiques sur une carte génomique construite avec GenVision (DNASTAR)

A) Pour *E. coli* K-12, les îlots correspondent aux gènes dont le type de produit est extrachromosomal (type *h*) selon la classification de *GenProtEC* [Serres *et al.*, 2004]. Si la valeur de GC_3 est inférieure à 43,6, elle est indiquée en bleu, si elle est supérieure à 57,9, elle est indiquée en rouge, et sinon en jaune.

B) Pour *P. luminescens*, les îlots annotés ont été représentés (*Article V* p. 332). Si la valeur de GC_3 est inférieure à 33,3, elle est indiquée en bleu, si elle est supérieure à 52,3, elle est indiquée en rouge, et sinon en jaune.

C) Pour *Y. pestis* CO92, les îlots annotés ont été représentés [Parkhill *et al.*, 2001b]. Si la valeur de GC_3 est inférieure à 40,2, elle est indiquée en bleu, si elle est supérieure à 55,0, elle est indiquée en rouge et sinon en jaune.

- Il possède des régions répétées (en cyan dans la phase 0, répétitions longues de type *distant*), notamment dans les régions d'intégrase et de transposase.
- Il possède des gènes d'usage des codons synonymes atypique (classe III), certains étant difficiles à prédire avec la matrice I (*e.g.* YPO1096, YPO1097).
- Il contient des décalages du cadre de lecture (*e.g.* YPO1084, YPO1086a).
- Les protéines encodées sont similaires à des protéines de phage, de transposon ou à des protéines de fonction inconnue.

Il existe trois groupes de synténie entre l'îlot de *Y. pestis* CO92 et les gènes de *P. luminescens* :

1. plu1064 et plu1065 (en marron) font partie de l'îlot génomique plu0958–plu1166 annoté chez *P. luminescens*, intervenant dans la biosynthèse des pili de type IV (*Article V* p. 332).
2. plu4470 et plu4471 (en brun) font partie de l'îlot plu4451–plu4477 annoté comme ADN de prophage chez *P. luminescens* (Fig. 11.4 p. 348).
3. plu1836 et plu1837 (en vert) font partie de l'îlot plu1820–plu1839 annoté comme ADN de prophage putatif (vestige, en jaune sur la figure 11.3 B p. 346).

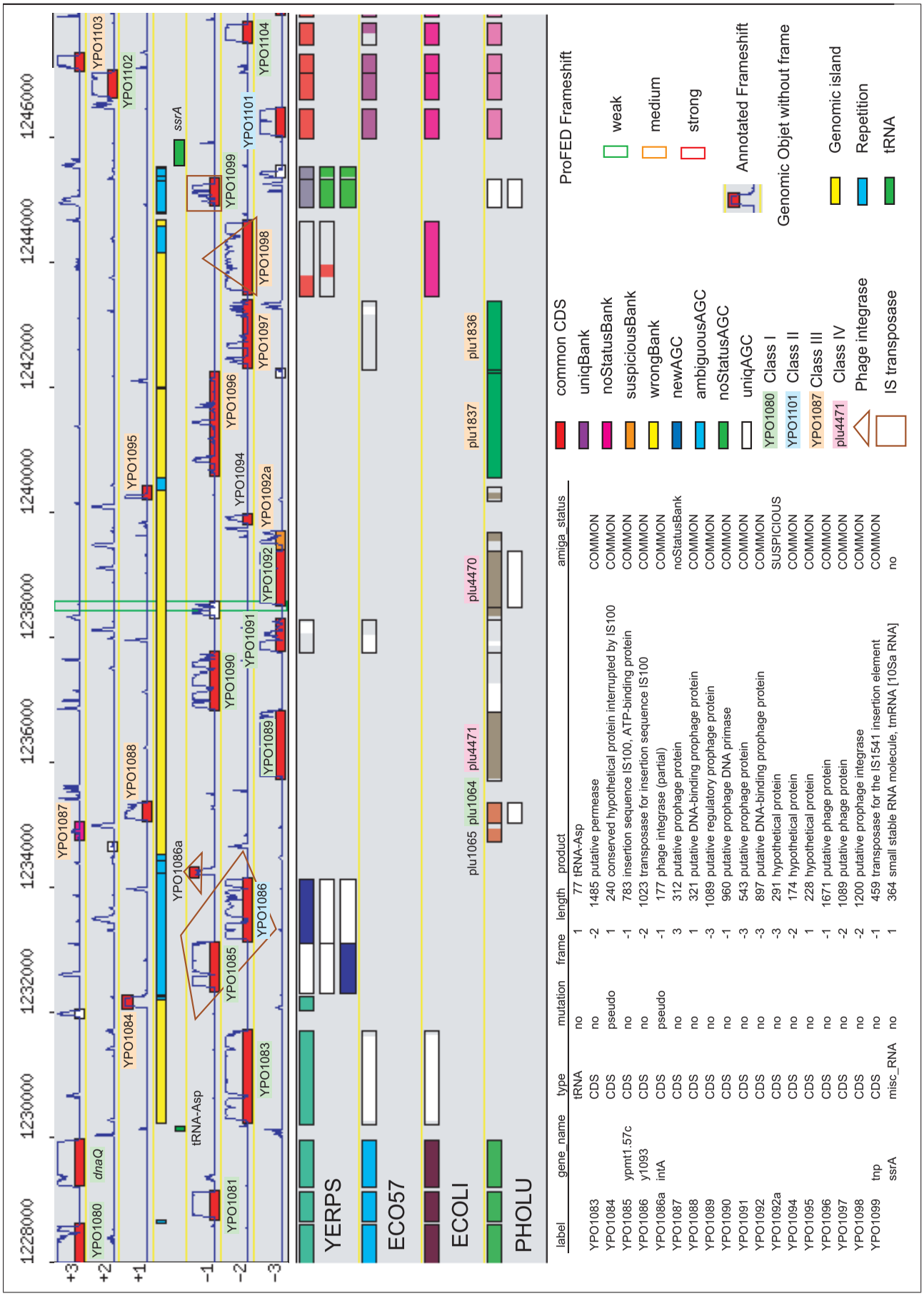
D'après les résultats de synténie présentés sur la figure 11.3 B p. 346, l'îlot plu1820 – plu1839 est spécifique de *P. luminescens* relativement aux quatre autres génomes d'entérobactéries. Il possède aussi des caractéristiques classiques d'un îlot génomique :

- Il est borné en $5'$ par une IS (plu1820) et en $3'$ par un gène d'ARNt.
- Il contient à la fois des gènes de transposase d'IS (*e.g.* plu1820, plu1824, plu1825) et un pseudogène d'intégrase de phage (*e.g.* plu1838, plu1839).
- Il possède des régions répétées (en cyan dans la phase 0), notamment dans la région de transposase.
- Il possède des gènes d'usage des codons synonymes atypique (classe III), certains étant difficiles à prédire avec la matrice I (*e.g.* plu1834, plu1835, plu1836, plu1837)
- Il contient des décalages du cadre de lecture (*e.g.* plu1825, plu1838, plu1839).
- Les protéines encodées sont similaires à des protéines de phage, de transposon ou à des protéines de fonction inconnue.

De plus, le processus de réannotation (voir p. 275) permet de prédire deux nouveaux gènes. Le produit de plu1824.1 ne présente pas de similitude dans UniProt. Le produit de plu1825.1 est similaire à une protéine hypothétique d'*E. coli* O6 (75% d'identité avec Q8FGH6). Cette CDS est qualifiée de '*noStatusAGC*' car elle chevauche plu1826 qui a une probabilité moyenne de codage de 0,47. Cependant, plu1826 est une protéine hypothétique qui ne présente pas de similitude dans les banques et qui pourraient donc être un faux-positif.

Ainsi, les îlots seraient préférentiellement au voisinage des ARNt, des intégrases et des transposases. Ils contiendraient une proportion plus importante de gènes mutés, de gènes de contenu en GC atypique (A+T riche ou G+C riche avec une dominance A+T riche) et de gènes d'usage des codons synonymes atypiques (*e.g.* classe III). Les classes II et III ne seraient pas exclusivement constituées respectivement de gènes hautement exprimés et de gènes HGT. Les gènes HGT ne se

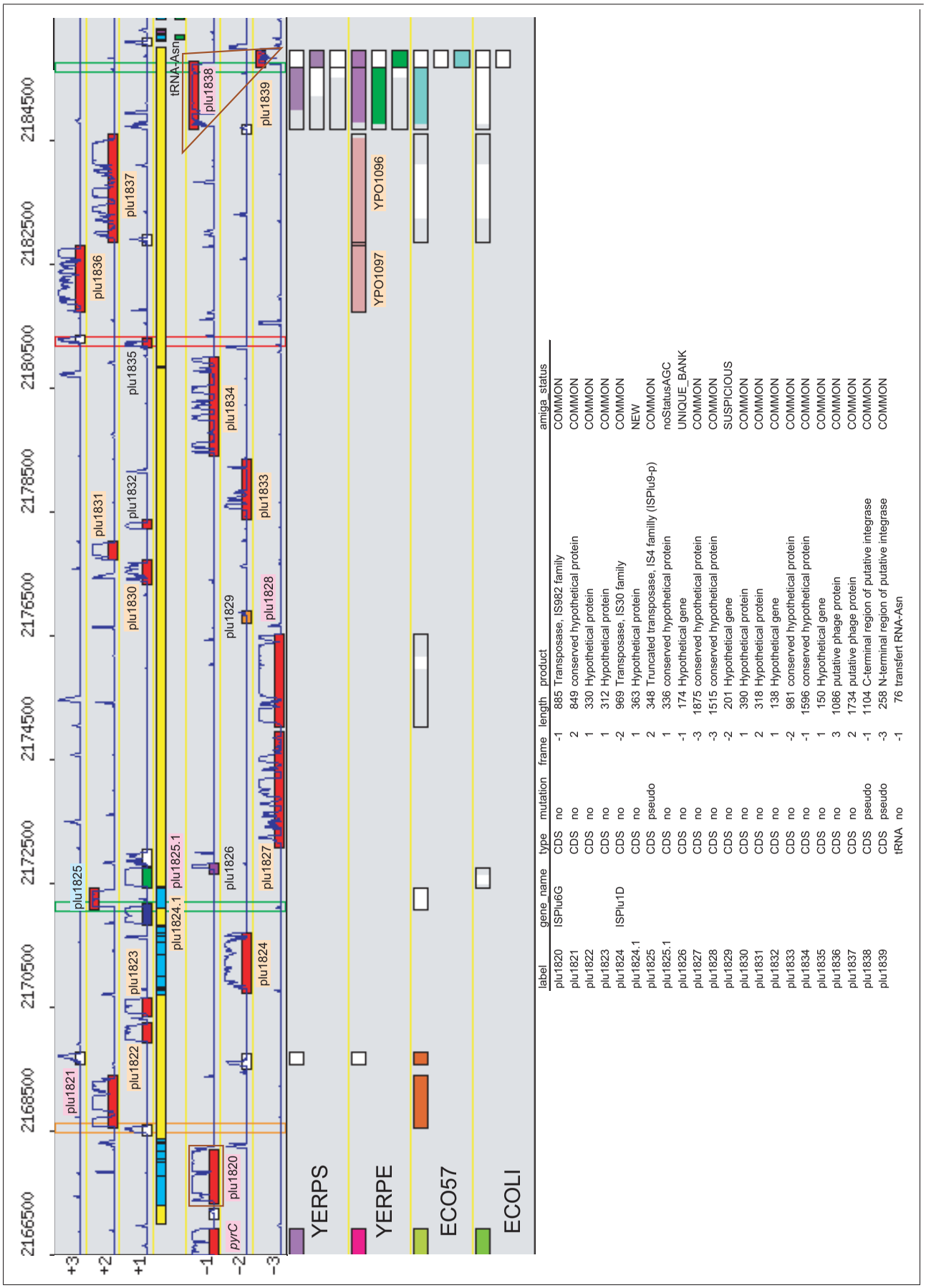
retrouveraient pas tous en classe III.
Ces hypothèses permettent d'affiner notre interprétation de l'effet Guttman : mathématiquement, les deux premiers axes de l'AFC sont non corrélés, mais biologiquement, ceci peut s'expliquer par le fait que les deux premiers axes sont liés par deux corrélations opposées et de même force (à l'origine de la forme de fer à cheval du nuage de gènes ; Fig. 11.1 p. 339). Cette interprétation reste à démontrer par des approches mixtes (intrinsèques, extrinsèques et expérimentales) afin de trouver un consensus sur les îlots génomiques.
Alors seulement, nous pourrons essayer de répondre à la question : « Comment reconnaître un îlot de pathogénie parmi les îlots génomiques ? ».



label	gene_name	type	mutation	frame	length	product	amiga_status
		tRNA	no	1	77	tRNA-Asp	
YPO1083		CDS	no	-2	1485	putative permease	COMMON
YPO1084		CDS	pseudo	1	240	conserved hypothetical protein interrupted by IS100	COMMON
YPO1085	ypmt1.57c	CDS	no	-1	783	insertion sequence IS100, ATP-binding protein	COMMON
YPO1086	y1093	CDS	no	-2	1023	transposase for insertion sequence IS100	COMMON
YPO1086a	intA	CDS	pseudo	-1	177	phage integrase (partial)	COMMON
YPO1087		CDS	no	3	312	putative prophage protein	noStatusBank
YPO1088		CDS	no	1	321	putative DNA-binding prophage protein	COMMON
YPO1089		CDS	no	-3	1089	putative regulatory prophage protein	COMMON
YPO1090		CDS	no	-1	960	putative prophage DNA primase	COMMON
YPO1091		CDS	no	-3	543	putative prophage protein	COMMON
YPO1092		CDS	no	-3	897	putative DNA-binding prophage protein	COMMON
YPO1092a		CDS	no	-3	291	hypothetical protein	SUSPICIOUS
YPO1094		CDS	no	-2	174	hypothetical protein	COMMON
YPO1095		CDS	no	1	228	hypothetical protein	COMMON
YPO1096		CDS	no	-1	1671	putative phage protein	COMMON
YPO1097		CDS	no	-2	1089	putative phage protein	COMMON
YPO1098		CDS	no	-2	1200	putative prophage integrase	COMMON
YPO1099	tnp	CDS	no	-1	459	transposase for the IS1541 insertion element	COMMON
	ssrA	misc_RNA	no	1	364	small stable RNA molecule, tmRNA [10Sa RNA]	no

FIG. 11.3 – A) Ilot de pathogénie chez *Y. pestis* CO92 annoté comme un vestige de phage [Parkhill *et al.*, 2001b]

La première carte de l'interface *MaGe* représente une portion du chromosome pivot. La seconde carte représente les synténies prédites entre les CDS du génome pivot et celles d'autres génomes [Labarre & Médigue, 2004]. Un rectangle blanc représente un orthologue isolé (absence de synténie). Un rectangle de couleur représente un orthologue colocalisé (présence de synténie). La partie transparente d'un rectangle blanc ou de couleur représente la partie de la protéine sujette qui ne s'aligne pas avec la protéine requête du génome pivot.



label	gene_name	type	mutation	frame	length	product	amiga_status
plu1820	ISPlu6G	CDS	no	-1	885	Transposase, IS982 family	COMMON
plu1821		CDS	no	2	849	conserved hypothetical protein	COMMON
plu1822		CDS	no	1	330	Hypothetical protein	COMMON
plu1823		CDS	no	1	312	Hypothetical protein	COMMON
plu1824	ISPlu1D	CDS	no	-2	969	Transposase, IS30 family	COMMON
plu1824.1		CDS	no	1	363	Hypothetical protein	NEW
plu1825		CDS	pseudo	2	348	Truncated transposase, IS4 familly (ISPlu9-p)	COMMON
plu1825.1		CDS	no	1	336	conserved hypothetical protein	noStatusAGC
plu1826		CDS	no	-1	174	Hypothetical gene	UNIQUE_BANK
plu1827		CDS	no	-3	1875	conserved hypothetical protein	COMMON
plu1828		CDS	no	-3	1515	conserved hypothetical protein	COMMON
plu1829		CDS	no	-2	201	Hypothetical gene	SUSPICIOUS
plu1830		CDS	no	1	390	Hypothetical protein	COMMON
plu1831		CDS	no	2	318	Hypothetical protein	COMMON
plu1832		CDS	no	1	138	Hypothetical gene	COMMON
plu1833		CDS	no	-2	981	conserved hypothetical protein	COMMON
plu1834		CDS	no	-1	1596	conserved hypothetical protein	COMMON
plu1835		CDS	no	1	150	Hypothetical gene	COMMON
plu1836		CDS	no	3	1086	putative phage protein	COMMON
plu1837		CDS	no	2	1734	putative phage protein	COMMON
plu1838		CDS	pseudo	-1	1104	C-terminal region of putative integrase	COMMON
plu1839		CDS	pseudo	-3	258	N-terminal region of putative integrase	COMMON
		tRNA	no	-1	76	transfert RNA-Asn	

FIG. 11.3 – B) Ilot génomique chez *P. luminescens* annoté comme un vestige de prophage (*Article V* p. 332)

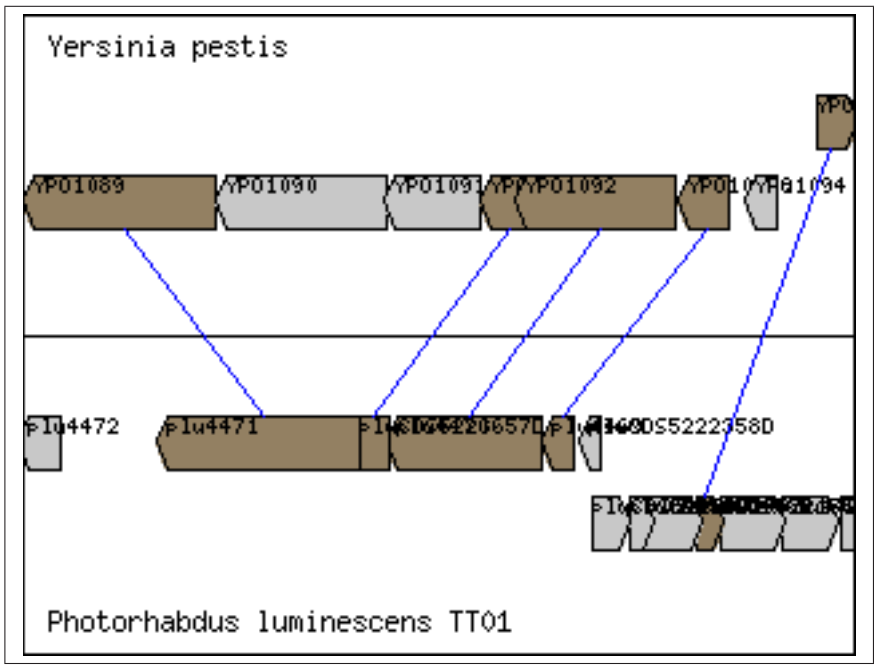


FIG. 11.4 – Visualisation du groupe de synténie YPO1089/plu4471–YPO1092/plu4470

Conclusion

Conclusions sur la (ré)annotation et perspectives d'exploration des génomes procaryotes complets

Conclusions

Aujourd'hui, l'Atelier de Génomique Comparative dispose d'un ensemble d'outils bioinformatiques pour modéliser, (ré)annoter, comparer, représenter, explorer des objets biologiques de génomes procaryotes complets.

Le modèle générique multigénome procaryote, PkGDB, permet de répondre à différents objectifs : annotations syntaxique, fonctionnelle et relationnelle d'un nouveau génome, réannotation de génomes publics. Une instance de PkGDB (*i.e.* base relationnelle) permet donc d'intégrer non seulement des ressources hétérogènes et distribuées (*e.g.* annotations brutes nucléiques, protéiques), mais aussi les résultats de diverses méthodes d'analyse de séquences (*e.g.* prédictions nucléiques, protéiques).

Nous avons aussi développé six stratégies dans le cadre d'un processus de réannotation :

1. PkGDB_Tools regroupe un ensemble de scripts pour l'analyse grammaticale d'un fichier d'annotations, le formatage d'annotations des banques ou de prédictions automatiques en objets génomiques, le chargement de ces objets génomiques dans la base PkGDB, la vérification de la cohérence des objets génomiques de PkGDB, la sélection et le le formatage d'objets génomiques de PkGDB, dans des standards tels que *csv* (chargement des objets dans le module *GenoBool*), *gbk* (chargement des objets dans le module *GenoAnnot* ou soumission des annotations à GenBank).
2. *AMIMat* est une stratégie experte semi-automatique d'apprentissage de séquences codantes et non-codantes (première phase de la stratégie globale de prédiction de CDS). Il s'agit de construire k matrices de transition en fonction des k classes de gènes d'usage des codons synonymes.
3. *AMIGene* est un programme automatique de reconnaissance et de post-traitement des CDS (seconde phase de la stratégie globale de prédiction de CDS). Il s'agit de calculer la meilleure

probabilité moyenne de codage des CDS en fonction des k matrices puis de les filtrer selon différents critères (l'utilisateur doit choisir un jeu de paramètres de filtrage). Il existe une variante permettant de calculer la meilleure probabilité moyenne de codage d'une liste de CDS (*e.g.* CDS annotées).

4. *CompAnnot* permet de comparer un jeu de CDS prédites par *AMIGene* et un jeu de CDS annotées dans les banques, correspondant à la même version d'une séquence chromosomique, et de définir ainsi trois listes : (i) les CDS communes aux banques et à *AMIGene*, (ii) les CDS uniques aux banques et (iii) les CDS uniques à *AMIGene*.

5. D'autres programmes permettent de rechercher des similitudes entre le produit des CDS et une banque de séquences protéiques, et d'attribuer un statut de similitude en fonction de l'analyse des résultats, ce qui nécessite une certaine expertise.

6. Le programme *SWAN* permet d'attribuer un statut de réannotation à certaines des CDS uniques (jeu de paramètres experts).

Nous avons ainsi intégré différentes approches (intrinsèque–extrinsèque) dans un processus expert de réannotation semi-automatique : analyse automatique et manuelle des CDS des banques, prédictions semi-automatiques, réconciliation automatique des annotations et des prédictions, attribution automatique de statuts de similitude et de réannotation.

Nous disposons essentiellement deux plates-formes : *Genostar* et *MaGe*. Nous avons commencé à intégrer nos ressources et nos stratégies de (ré)annotation dans la plate-forme d'exploration d'objets génomiques, *Genostar*. D. Vallenet a développé une interface de (ré)annotation et d'exploration au dessus de la base PkGDB : *MaGe*. Elle peut être branchée sur différentes instances du modèle générique PkGDB, par exemple :

- *Acineto*DB pour l'annotation manuelle du nouveau génome d'*Acinetobacter* ADP1 [Barbe *et al.*, 2004] à la lumière des résultats de prédictions automatiques, comme les synténies ou les voies métaboliques.
- *Neisseria*DB pour la comparaison des génomes des *Neisseria* à la lumière des résultats de prédictions, de réannotations, de synténies.
- *Entero*DB pour l'exploration des îlots génomiques d'entérobactéries à la lumière des résultats de réannotations, de synténies.

D. Vallenet a aussi développé une variante de *MaGe*, CompAnnotViewer, pour la correction manuelle et experte des bornes anormales des CDS des banques. Il a enfin mis en ligne une application d'*AMIGene*, disponibles sur le serveur Web du Genoscope.

L'ensemble de ces trois composants : base de données (*e.g.* PkGDB), méthodes d'analyse de séquences (*e.g. AMIGene*) et interface de représentation cartographique, synthétique et dynamique des résultats (*e.g. MaGe*), peut être considéré comme une plate-forme d'exploration des annotations des génomes procaryotes.

Perspectives

A terme, nous souhaiterions que le processus de réannotation soit intégré dans cinq structures complémentaires, capables de communiquer :

- PkGDB et *Genostar* (*e.g.* stratégie semi-automatique de prédiction de CDS d'un chromosome),
- PkGDB et BioFacet pour la recherche de similitude entre deux séquences à grande échelle,
- PkGDB et *MaGe* pour l'exploration et la validation manuelle des annotations à la lumière des groupes de synténie.
- PkGDB et une plate-forme de validation expérimentale (*e.g.* puces à ADN, caractérisations phénotypiques et complémentations de banques de mutants) pour confirmer et enrichir des annotations.

L'ensemble de ces cinq structures peut être exploité selon différents objectifs, par exemple, des projets d'annotation et d'exploration de génomes procaryotes nouvellement séquencés au sein d'un centre de séquençage comme le Genoscope, ou des projets de réannotation comme le projet *HAMAP* dédié aux protéomes microbiens publics complets.

Aussi, nous possédons tous les outils nécessaires au développement d'une stratégie de prédiction d'îlots génomiques [Bocs *et al.*, 2001]. Il suffirait de combiner plusieurs critères comme la présence d'ARNt, de gènes de mobilité, de décalages du cadre de lecture, de répétitions, le pourcentage en GC_3 des CDS, leur classe d'usage des codons synonymes, la recherche de mots-clés (*e.g.* virulence, résistance aux stress (antibiotique), inconnues (ORFan)), la recherche de similitudes, etc. Ce projet innovant m'a permis d'être sélectionnée au concours Anvar 2002 (Fig. 11.5 p. 354) :

« *Assurée de l'importance des processus transversaux dans le cadre du séquençage des génomes à grande échelle, la société aura pour vocation d'accélérer le développement d'applications dans les domaines-clés de la biopharmacie, de l'épidémiologie, de l'agroalimentaire ou de l'environnement.*

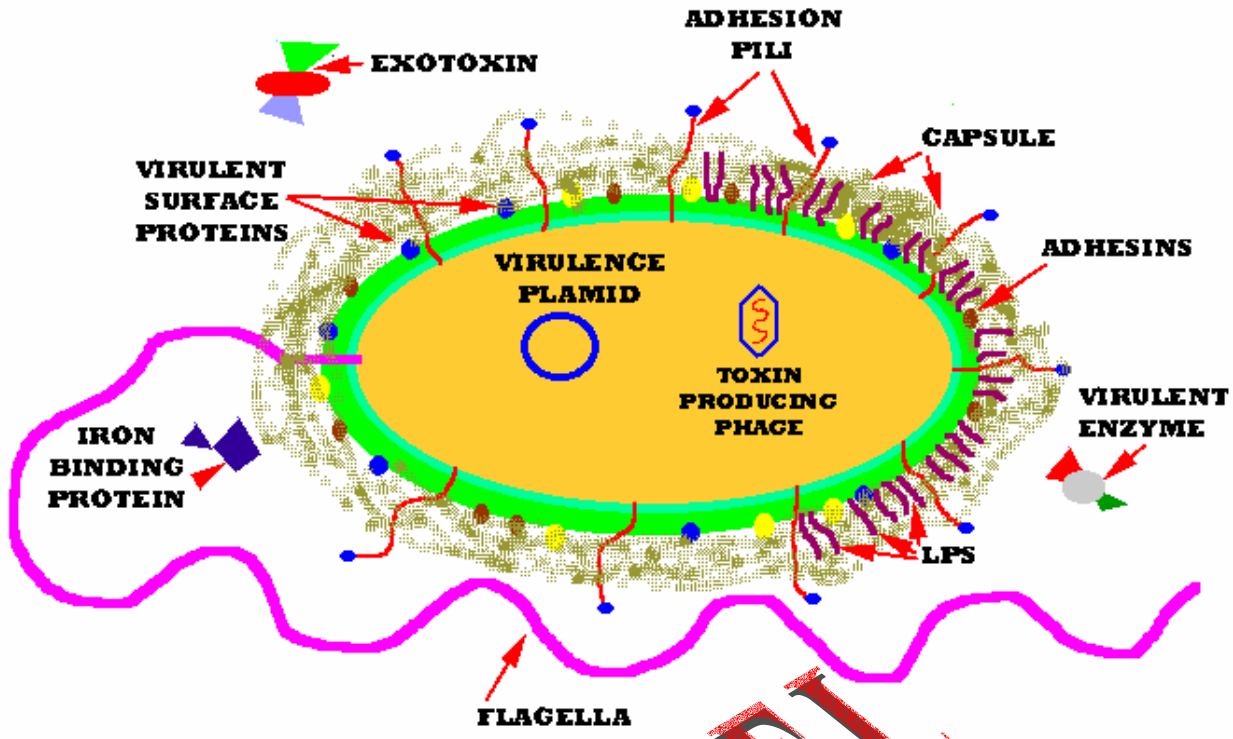
Force est de constater qu'il existe un vide à combler entre la nouvelle science qu'est la bioinformatique et les biotechnologies. Concrètement la première fonction de la société sera d'utiliser un logiciel pour prédire des régions génomiques impliquées dans les mécanismes de virulence des bactéries pathogènes et de démontrer expérimentalement la fonction des protéines prédites. Ainsi la liste de protéines de virulence d'une bactérie pathogène aidera à trouver des solutions contre la maladie infectieuse provoquée par cet organisme. »

Ce projet permettrait, d'un point de vue appliqué, de breveter des îlots de pathogénie dans le but de découvrir de nouvelles classes d'antibiotiques et d'un point de vue fondamental, de comprendre le rôle du transfert horizontal de gènes dans l'évolution des microbes afin de reconstruire une histoire la plus réaliste possible.

Génomes bactériens :

Une base dans l'ordinateur,

Une banque au congélateur.



Facteurs de virulence bactériens

Fiche d'identité MicobIG :
Lieu : Genopole Evry (Essonne)
Spécialité : Biotechnologie, bioinformatique, microbiologie
Création : Janvier 2003
Statut : SARL (350 000 euros)
Contact : sbocs@genoscope.cns.fr
06-84-97-85-30

FIG. 11.5 – Page de garde du projet « en émergence », génomes bactériens : une base dans l'ordinateur et une banque au congélateur, présenté au concours Création d'Entreprises de Technologie Innovantes 2002 (4ème édition)

Bibliographie

Bibliographie

[Achaz *et al.*, 2002] Achaz, G., Rocha, E. P., Netter, P., and Coissac, E. (2002). "Origin and fate of repeats in bacteria". Nucleic Acids Res 30(13) :2987–94.

[Achtman *et al.*, 1999] Achtman, M. *et al.* (1999). "Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis". Proc Natl Acad Sci U S A 96(24) :14043–8.

[Allen *et al.*, 2004] Allen, J. E., Pertea, M., and Salzberg, S. L. (2004). "Computational gene prediction using multiple sources of evidence". Genome Res 14(1) :142–8.

[Altschul, 1991] Altschul, S. F. (1991). "Amino acid substitution matrices from an information theoretic perspective". J Mol Biol 219(3) :555–65.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). "Basic local alignment search tool". J Mol Biol 215(3) :403–10.

[Altschul *et al.*, 1997] Altschul, S. F. *et al.* (1997). "Gapped BLAST and PSI-BLAST : a new generation of protein database search programs". Nucleic Acids Res 25(17) :3389–402.

[Amann, 2002] Amann, R. (2002). "Think big : the international dimension of environmental microbiology". Environ Microbiol 4(1) :3.

[Amann *et al.*, 1995] Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). "Phylogenetic identification and in situ detection of individual microbial cells without cultivation". Microbiol Rev 59(1) :143–69.

[Andrade *et al.*, 1997] Andrade, M. *et al.* (1997). "Sequence analysis of the Methanococcus jannaschii genome and the prediction of protein function". Comput Appl Biosci 13(4) :481–3.

[Andrade *et al.*, 1999] Andrade, M. A. *et al.* (1999). "Automated genome sequence analysis and annotation". Bioinformatics 15(5) :391–412.

[Apweiler *et al.*, 2004] Apweiler, R. *et al.* (2004). "UniProt : the Universal Protein knowledgebase". Nucleic Acids Res 32 Database issue :D115–9.

[Ashburner *et al.*, 2000] Ashburner, M. *et al.* (2000). "Gene ontology : tool for the unification of biology. The Gene Ontology Consortium". Nat Genet 25(1) :25–9.

[Audic & Claverie, 1998] Audic, S., and Claverie, J. M. (1998). "Self-identification of protein-coding regions in microbial genomes". Proc Natl Acad Sci U S A 95(17) :10026–31.

[Avery *et al.*, 1944] Avery, O., MacLeod, C., and M., M. (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III.". The Journal of Experimental Medecine 79 :137–159.

[Azad & Borodovsky, 2004] Azad, R. K., and Borodovsky, M. (2004). "Effects of choice of DNA sequence model structure on gene identification accuracy". Bioinformatics.

[Baccini & Besse, 2002] Baccini, A., and Besse, P. (2002). "Data mining 1. Exploration Statistique". Supports de cours présentant les techniques usuelles de description ou modélisation statistique et de data mining.

[Bader *et al.*, 2003] Bader, G. D., Betel, D., and Hogue, C. W. (2003). "BIND : the Biomolecular Interaction Network Database". Nucleic Acids Res 31(1) :248–50.

[Bairoch, 2000] Bairoch, A. (2000). "The ENZYME database in 2000". Nucleic Acids Res 28(1) :304–5.

[Barbe *et al.*, 2004] Barbe, V. *et al.* (2004). "Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium". Nucleic Acids Res 32(19) :5766–79.

[Baxevanis, 2003] Baxevanis, A. D. (2003). "The Molecular Biology Database Collection : 2003 update". Nucleic Acids Res 31(1) :1–12.

[Bejerano, 2003] Bejerano, G. (2003). "Efficient Exact p-Value Computation and Applications to Biosequence Analysis.". In Miller, W., Vingron, M., Istrail, S., Pevzner, P., and Waterman, M., editors, *RECOMB 2003 Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, pages 38–47.

[Bennetzen & Hall, 1982] Bennetzen, J. L., and Hall, B. D. (1982). "Codon selection in yeast". J Biol Chem 257(6) :3026–31.

[Benson *et al.*, 2004] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). "GenBank : update". Nucleic Acids Res 32 Database issue :D23–6.

[Bentz, 2000] Bentz, C. (2000). "Développement d'une base de données dédiées aux génomes pathogènes.". Ingénieur informaticien première année, CNAM Institut d'Informatique d'Entreprise.

[Benzécri, 1973] Benzécri, J. (1973). "L'analyse des données.". Paris, Dunod.

[Bernal *et al.*, 2001] Bernal, A., Ear, U., and Kyrpides, N. (2001). "Genomes OnLine Database (GOLD) : a monitor of genome projects world-wide". Nucleic Acids Res 29(1) :126–7.

[Bernstein *et al.*, 1977] Bernstein, F. C. *et al.* (1977). "The Protein Data Bank : a computer-based archival file for macromolecular structures". J Mol Biol 112(3) :535–42.

[Besemer & Borodovsky, 1999] Besemer, J., and Borodovsky, M. (1999). "Heuristic approach to deriving models for gene finding". Nucleic Acids Res 27(19) :3911–20.

[Besemer *et al.*, 2001] Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). "GeneMarkS : a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions". Nucleic Acids Res 29(12) :2607–18.

[Biaudet *et al.*, 1997] Biaudet, V., Samson, F., and Bessières, P. (1997). "Micado–a network-oriented database for microbial genomes". Comput Appl Biosci 13(4) :431–8.

[Billoud *et al.*, 1996] Billoud, B., Kontic, M., and Viari, A. (1996). "Palingol : a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database". Nucleic Acids Res 24(8) :1395–403.

[Blanchette & Tompa, 2003] Blanchette, M., and Tompa, M. (2003). "FootPrinter : a program designed for phylogenetic footprinting". Nucleic Acids Res 31(13) :3840–2.

[Blattner *et al.*, 1997] Blattner, F. R. *et al.* (1997). "The complete genome sequence of Escherichia coli K-12". Science 277(5331) :1453–74.

[Bocs, 1999] Bocs, S. (1999). "Intégration de la stratégie Proscan/Prosite dans l'environnement Imagene.". Dess d'informatique appliquée à la biologie, Université Paris VI - Pierre et Marie Curie.

[Bocs, 2000] Bocs, S. (2000). "Détection automatique de phases codantes dans les génomes procaryotes.". Colloque Traitement et Analyse des séquences (action ministérielle Informatique, Mathématique et Physique pour la Génomique (IMPG)), Evry (France).

[Bocs *et al.*, 2000] Bocs, S., Blazy, S., Glaser, P., and Médigue, C. (2000). "An automatic detection of prokaryotic CoDing Sequence combining several independant methods.". Genomes 2000 : International Conference on Microbial and Model Genomes (American Society for Microbiology (ASM) and Institut Pasteur (IP)), Paris (France).

[Bocs *et al.*, 2003] Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Médigue, C. (2003). "AMIGene : Annotation of MIcrobial Genes". Nucleic Acids Res 31(13) :3723–6.

[Bocs *et al.*, 2002] Bocs, S., Danchin, A., and Médigue, C. (2002). "Re-annotation of genome microbial CoDing-Sequences : finding new genes and inaccurately annotated genes". BMC Bioinformatics 3(1) :5.

[Bocs *et al.*, 2001] Bocs, S., Nicolas, P., Devine, C., Glaser, P., and Médigue, C. (2001). "Searching for virulence factors on pathogenic enterobacteria genomic sequences.". 5th Computational Genomics Conference (The Institute for Genomic Research (TIGR)), Baltimore MD (USA).

[Boeckmann *et al.*, 2003] Boeckmann, B. *et al.* (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003". Nucleic Acids Res 31(1) :365–70.

[Boneca *et al.*, 2003] Boneca, I. G. *et al.* (2003). "A revised annotation and comparative analysis of Helicobacter pylori genomes". Nucleic Acids Res 31(6) :1704–14.

[Borodovsky & McIninch, 1993a] Borodovsky, M., and McIninch, J. (1993a). "GENMARK : Parallel gene recognition for both DNA strands.". Computers Chem. 17 :123–133.

[Borodovsky & McIninch, 1993b] Borodovsky, M., and McIninch, J. (1993b). "Prediction of genes locations using DNA Markov chain models.". In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, pages 231–248.

[Borodovsky *et al.*, 1995] Borodovsky, M. *et al.* (1995). "Detection of new genes in a bacterial genome using Markov models for three gene classes". Nucleic Acids Res 23(17) :3554–62.

[Bourne *et al.*, 2004] Bourne, P. E. *et al.* (2004). "The distribution and query systems of the RCSB Protein Data Bank". Nucleic Acids Res 32 Database issue :D223–5.

[Bouroche & Saporta, 1992] Bouroche, J.-M., and Saporta, G. (1992). "QUE SAIS-JE ? L'analyse des données.". Presses Universitaires de France.

[Brown *et al.*, 1998] Brown, N. P., Sander, C., and Bork, P. (1998). "Frame : detection of genomic sequencing errors". Bioinformatics 14(4) :367–71.

[Bulmer, 1991] Bulmer, M. (1991). "The selection-mutation-drift theory of synonymous codon usage". Genetics 129(3) :897–907.

[Bult *et al.*, 1996] Bult, C. J. *et al.* (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii". Science 273(5278) :1058–73.

[Camus *et al.*, 2002] Camus, J. C., Pryor, M. J., Médigue, C., and Cole, S. T. (2002). "Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv". Microbiology 148(Pt 10) :2967–73.

[Carbone *et al.*, 2003] Carbone, A., Zinovyev, A., and Kepes, F. (2003). "Codon adaptation index as a measure of dominating codon bias". Bioinformatics 19(16) :2005–15.

[Chain *et al.*, 2004] Chain, P. S. *et al.* (2004). "Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis". Proc Natl Acad Sci U S A 101(38) :13826–31.

[Chambaud, 2000] Chambaud, I. (2000). "Séquençage et analyse du génome de *Mycoplasma pulmonis*, l'agent étiologique de la Mycoplasmose Respiratoire Murine.". Biochimie – biologie moléculaire et cellulaire, Ecole Nationnale Supérieure Agronomique de Rennes.

[Chambaud *et al.*, 2001] Chambaud, I. *et al.* (2001). "The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis". Nucleic Acids Res 29(10) :2145–53.

[Chargaff, 1950] Chargaff, E. (1950). "Chemical specificity of nucleic acids and the mechanism of their enzymatic degradation.". Experientia 6 :201–209.

[Charlebois *et al.*, 2003] Charlebois, R. L., Beiko, R. G., and Ragan, M. A. (2003). "Microbial phylogenomics : Branching out". Nature 421(6920) :217.

[Charras & Lecroq,] Charras, C., and Lecroq, T. "Handbook of Exact String-Matching Algorithms". 30 algorithmes de recherche exacte d'un mot.

[Chen *et al.*, 1995] Chen, Q. K., Hertz, G. Z., and Stormo, G. D. (1995). "MATRIX SEARCH 1.0 : a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices". Comput Appl Biosci 11(5) :563–6.

[Chiapello, 1999] Chiapello, H. (1999). “Analyse comparée de l’usage des codons dans les génomes de plantes.”. Physiologie cellulaire et moléculaire des plantes, Université Paris VI - Pierre et Marie Curie.

[Chung & Wong, 1999] Chung, S. Y., and Wong, L. (1999). “Kleisli : a new tool for data integration in biology”. Trends Biotechnol 17(9) :351–5.

[Claudel-Renard, 2003] Claudel-Renard, C. (2003). “Inférence fonctionnelle et prédiction de voies métaboliques. Application à la bactérie fixatrice d’azote *Sinorhizobium meliloti*.”. Ecole doctorale biologie santé biotechnologies (bioinformatique), Université Paul Sabatier Toulouse III.

[Claudel-Renard *et al.*, 2003] Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). “Enzyme-specific profiles for genome annotation : PRIAM”. Nucleic Acids Res 31(22) :6633–9.

[Claverie & States, 1993] Claverie, J.-M., and States, D. J. (1993). “Information enhancement methods for large scale sequence analysis.”. Comput. Chem. 17 :191–201.

[Cohan, 2002] Cohan, F. M. (2002). “What are bacterial species ?”. Annu Rev Microbiol 56 :457–87.

[Cole *et al.*, 2003] Cole, J. R. *et al.* (2003). “The Ribosomal Database Project (RDP-II) : previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy”. Nucleic Acids Res 31(1) :442–3.

[Cole *et al.*, 1998] Cole, S. T. *et al.* (1998). “Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence”. Nature 393(6685) :537–44.

[Craig, 1996] Craig, N. L. (1996). “Transposition.”. In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 124. American Society for Microbiology Press.

[Croquette & Carlier, 1999] Croquette, A., and Carlier, A. (1999). “Classification Automatique”. Support de cours présentant le chapitre 6 d’Analyse des Données Multidimensionnelles.

[Cruveiller *et al.*, 2003a] Cruveiller, S. *et al.* (2003a). “Gene Classes in Bacterial Genomes.”. Proceedings of the European Conference on Computational Biology (ECCB’2003), Paris (France).

[Cruveiller *et al.*, 2003b] Cruveiller, S., Jabbari, K., Clay, O., and Bemardi, G. (2003b). “Compositional features of eukaryotic genomes for checking predicted genes”. Brief Bioinform 4(1) :43–52.

[Danchin, 1998] Danchin, A. (1998). “LA BARQUE DE DELPHES ce que révèle le texte des génomes”. ODILE JACOB.

[Danchin & Sekowska, 1993] Danchin, A., and Sekowska, A. (1993). “*Bacillus subtilis*.”. In Sahm, H., editor, *Biotechnology*, volume 1, chapter 12. Wiley-VCH, Weinheim, second, completely revised edition edition.

[Dandekar *et al.*, 2000] Dandekar, T. *et al.* (2000). “Re-annotating the Mycoplasma pneumoniae genome sequence : adding value, function and reading frames”. Nucleic Acids Res 28(17) :3278–88.

[Dardel & Képès, 2002] Dardel, F., and Képès, F. (2002). “Statistiques et séquences.”. In *Bioinformatique, Génomique et post-génomique*, chapter 5. LES EDITIONS DE L’ECOLE POLYTECHNIQUE.

[d'Aubenton Carafa *et al.*, 1990] d'Aubenton Carafa, Y., Brody, E., and Thermes, C. (1990). "Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures". J Mol Biol 216(4) :835–58.

[Daubin *et al.*, 2003a] Daubin, V., Lerat, E., and Perrière, G. (2003a). "The source of laterally transferred genes in bacterial genomes". Genome Biol 4(9) :R57.

[Daubin *et al.*, 2003b] Daubin, V., Moran, N. A., and Ochman, H. (2003b). "Phylogenetics and the cohesion of bacterial genomes". Science 301(5634) :829–32.

[Dayhoff *et al.*, 1965] Dayhoff, M., Eck, R., Chang, M., and Sochard, M. (1965). "Atlas of Protein Sequence and Structure.", volume 1. National Biomedical Research Foundation, Silver Spring, MD.

[Delcher *et al.*, 1999a] Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999a). "Improved microbial gene identification with GLIMMER". Nucleic Acids Res 27(23) :4636–41.

[Delcher *et al.*, 1999b] Delcher, A. L. *et al.* (1999b). "Alignment of whole genomes". Nucleic Acids Res 27(11) :2369–76.

[Delorme & Henaut, 1988] Delorme, M. O., and Henaut, A. (1988). "Merging of distance matrices and classification by dynamic clustering". Comput Appl Biosci 4(4) :453–8.

[Den Rupp, 1996] Den Rupp, W. (1996). "DNA Repair Mechanism.". In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 118. American Society for Microbiology Press.

[Deschavanne & Filipski, 1995] Deschavanne, P., and Filipski, J. (1995). "Correlation of GC content with replication timing and repair mechanisms in weakly expressed E.coli genes". Nucleic Acids Res 23(8) :1350–3.

[Devine, 2001] Devine, C. (2001). "Interfaçage et mise en ligne de la base des génomes procaryotes : PkGDB.". Ingénieur informaticien première année, CNAM Institut d'Informatique d'Entreprise.

[Devos & Valencia, 2001] Devos, D., and Valencia, A. (2001). "Intrinsic errors in genome annotation". Trends Genet 17(8) :429–31.

[Diaz-Lazcoz *et al.*, 1995] Diaz-Lazcoz, Y., Henaut, A., Vigier, P., and Risler, J. L. (1995). "Differential codon usage for conserved amino acids : evidence that the serine codons TCN were primordial". J Mol Biol 250(2) :123–7.

[Diday, 1971] Diday, E. (1971). "Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques.". Revue de Staitisque Appliquée XIX :19–33.

[Discala *et al.*, 2000] Discala, C., Benigni, X., Barillot, E., and Vaysseix, G. (2000). "DBcat : a catalog of 500 biological databases". Nucleic Acids Res 28(1) :8–9.

[Duchaud *et al.*, 2003] Duchaud, E. *et al.* (2003). "The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens". Nat Biotechnol 21(11) :1307–13.

[Dundon *et al.*, 1999] Dundon, W. G., Marshall, D. G., Morain, C. A., and Smyth, C. J. (1999). "A novel tRNA-associated locus (trl) from Helicobacter pylori is co-transcribed with tRNA(Gly) and reveals genetic diversity". Microbiology 145 (Pt 6) :1289–98.

[Durand *et al.*, 2003] Durand, P. *et al.* (2003). "Integration of data and methods for genome analysis". Curr Opin Drug Discov Devel 6(3) :346–52.

[Durbin *et al.*, 2001a] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001a). "Biological sequence analysis.". Cambridge University Press.

[Durbin *et al.*, 2001b] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001b). "Introduction.". In *Biological sequence analysis*, chapter 1. Cambridge University Press.

[Durbin *et al.*, 2001c] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001c). "Making tree from pairwise distances.". In *Biological sequence analysis*, chapter 7. Cambridge University Press.

[Durbin *et al.*, 2001d] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001d). "Markov chains and hidden Markov models.". In *Biological sequence analysis*, chapter 3. Cambridge University Press.

[Durbin *et al.*, 2001e] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001e). "Pairwise alignement.". In *Biological sequence analysis*, chapter 2. Cambridge University Press.

[Durbin *et al.*, 2001f] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001f). "Profile HMM for sequence families.". In *Biological sequence analysis*, chapter 5. Cambridge University Press.

[Durbin *et al.*, 2001g] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001g). "RNA structure analysis.". In *Biological sequence analysis*, chapter 10. Cambridge University Press.

[Durbin *et al.*, 2001h] Durbin, R., Eddy, S., A., K., and Mitchison, G. (2001h). "Transformational grammars.". In *Biological sequence analysis*, chapter 9. Cambridge University Press.

[Eddy & Durbin, 1994] Eddy, S. R., and Durbin, R. (1994). "RNA sequence analysis using covariance models". Nucleic Acids Res 22(11) :2079–88.

[el Mabrouk & Lisacek, 1996] el Mabrouk, N., and Lisacek, F. (1996). "Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome". J Mol Biol 264(1) :46–55.

[Ermolaeva *et al.*, 2000] Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O., and Salzberg, S. L. (2000). "Prediction of transcription terminators in bacterial genomes". J Mol Biol 301(1) :27–33.

[Etzold *et al.*, 1996] Etzold, T., Ulyanov, A., and Argos, P. (1996). "SRS : information retrieval system for molecular biology data banks". Methods Enzymol 266 :114–28.

[ffrench Constant & Bowen, 2000] ffrench Constant, R. H., and Bowen, D. J. (2000). "Novel insecticidal toxins from nematode-symbiotic bacteria". Cell Mol Life Sci 57(5) :828–33.

[Fichant & Burks, 1991] Fichant, G. A., and Burks, C. (1991). "Identifying potential tRNA genes in genomic DNA sequences". J Mol Biol 220(3) :659–71.

[Fichant & Quentin, 1995] Fichant, G. A., and Quentin, Y. (1995). "A frameshift error detection algorithm for DNA sequencing projects". Nucleic Acids Res 23(15) :2900–8.

[Fickett, 1982] Fickett, J. W. (1982). "Recognition of protein coding regions in DNA sequences". Nucleic Acids Res 10(17) :5303–18.

[Fickett & Hatzigeorgiou, 1997] Fickett, J. W., and Hatzigeorgiou, A. G. (1997). "Eukaryotic promoter recognition". Genome Res 7(9) :861–78.

[Fickett & Tung, 1992] Fickett, J. W., and Tung, C. S. (1992). "Assessment of protein coding measures". Nucleic Acids Res 20(24) :6441–50.

[Finlay & Falkow, 1997] Finlay, B. B., and Falkow, S. (1997). "Common themes in microbial pathogenicity revisited". Microbiol Mol Biol Rev 61(2) :136–69.

[Fitch, 1970] Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins". Syst Zool 19(2) :99–113.

[Fitch, 2000] Fitch, W. M. (2000). "Homology a personal view on some of the problems". Trends Genet 16(5) :227–31.

[Fleischmann *et al.*, 2004] Fleischmann, A. *et al.* (2004). "IntEnz, the integrated relational enzyme database". Nucleic Acids Res 32 Database issue :D434–7.

[Fleischmann *et al.*, 1995] Fleischmann, R. D. *et al.* (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". Science 269(5223) :496–512.

[Fleischmann *et al.*, 2002] Fleischmann, R. D. *et al.* (2002). "Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains". J Bacteriol 184(19) :5479–90.

[Florea *et al.*, 2003] Florea, L., McClelland, M., Riemer, C., Schwartz, S., and Miller, W. (2003). "EnteriX 2003 : Visualization tools for genome alignments of Enterobacteriaceae". Nucleic Acids Res 31(13) :3527–32.

[Florea *et al.*, 2000] Florea, L. *et al.* (2000). "Web-based visualization tools for bacterial genome alignments". Nucleic Acids Res 28(18) :3486–96.

[Forgy, 1965] Forgy, E. (1965). "Cluster Analysis of multivariate data : Efficiency vs. interpretability of classifications.". Biometrics 21 :768.

[Frank & Lobry, 1999] Frank, A. C., and Lobry, J. R. (1999). "Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms". Gene 238(1) :65–77.

[Frank & Lobry, 2000] Frank, A. C., and Lobry, J. R. (2000). "Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes". Bioinformatics 16(6) :560–1.

[Friis *et al.*, 2000] Friis, C., Jensen, L. J., and Ussery, D. W. (2000). "Visualization of pathogenicity regions in bacteria". Genetica 108(1) :47–51.

[Frishman *et al.*, 1998a] Frishman, D., Heumann, K., Lesk, A., and Mewes, H. W. (1998a). "Comprehensive, comprehensible, distributed and intelligent databases : current status". Bioinformatics 14(7) :551–61.

[Frishman *et al.*, 1999] Frishman, D., Mironov, A., and Gelfand, M. (1999). "Starts of bacterial genes : estimating the reliability of computer predictions". Gene 234(2) :257–65.

[Frishman *et al.*, 1998b] Frishman, D., Mironov, A., Mewes, H. W., and Gelfand, M. (1998b). "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes". Nucleic Acids Res 26(12) :2941–7.

[Frishman *et al.*, 2003] Frishman, D. *et al.* (2003). "The PEDANT genome database". Nucleic Acids Res 31(1) :207–11.

[Gaasterland *et al.*, 2000] Gaasterland, T. *et al.* (2000). "MAGPIE/EGRET annotation of the 2.9-Mb Drosophila melanogaster Adh region". Genome Res 10(4) :502–10.

[Gaasterland & Sensen, 1996] Gaasterland, T., and Sensen, C. W. (1996). "MAGPIE : automated genome interpretation". Trends Genet 12(2) :76–8.

[Garcia-Vallve *et al.*, 2003] Garcia-Vallve, S., Guzman, E., Montero, M. A., and Romeu, A. (2003). "HGT-DB : a database of putative horizontally transferred genes in prokaryotic complete genomes". Nucleic Acids Res 31(1) :187–9.

[Gattiker *et al.*, 2003] Gattiker, A. *et al.* (2003). "Automated annotation of microbial proteomes in SWISS-PROT". Comput Biol Chem 27(1) :49–58.

[Ghosh, 1998] Ghosh, D. (1998). "OOTFD (Object-Oriented Transcription Factors Database) : an object-oriented successor to TFD". Nucleic Acids Res 26(1) :360–2.

[Glasner *et al.*, 2003] Glasner, J. D. *et al.* (2003). "ASAP, a systematic annotation package for community analysis of genomes". Nucleic Acids Res 31(1) :147–51.

[Glemet & Codani, 1997] Glemet, E., and Codani, J. J. (1997). "LASSAP, a LArge Scale Sequence compArison Package". Comput Appl Biosci 13(2) :137–43.

[Goffeau *et al.*, 1996] Goffeau, A. *et al.* (1996). "Life with 6000 genes". Science 274(5287) :546, 563–7.

[Gogarten *et al.*, 2002] Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). "Prokaryotic evolution in light of gene transfer". Mol Biol Evol 19(12) :2226–38.

[Gollub *et al.*, 2003] Gollub, J. *et al.* (2003). "The Stanford Microarray Database : data access and quality assessment tools". Nucleic Acids Res 31(1) :94–6.

[Gonnet *et al.*, 2000] Gonnet, G. H., Hallett, M. T., Korostensky, C., and Bernardin, L. (2000). "Darwin v. 2.0 : an interpreted computer language for the biosciences". Bioinformatics 16(2) :101–3.

[Goto *et al.*, 2002] Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002). "LIGAND : database of chemical compounds and reactions in biological pathways". Nucleic Acids Res 30(1) :402–4.

[Gouy & Gautier, 1982] Gouy, M., and Gautier, C. (1982). "Codon usage in bacteria : correlation with gene expressivity". Nucleic Acids Res 10(22) :7055–74.

[Grantham *et al.*, 1980] Grantham, R., Gautier, C., and Gouy, M. (1980). "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type". Nucleic Acids Res 8(9) :1893–912.

[Grantham *et al.*, 1981] Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). "Codon catalog usage is a genome strategy modulated for gene expressivity". Nucleic Acids Res 9(1) :r43–74.

[Greenacre, 1984] Greenacre, M. J. (1984). "Theory and Application of Correspondence Analysis.". New York Academic Press.

[Gribaldo & Philippe, 2002] Gribaldo, S., and Philippe, H. (2002). "Ancient phylogenetic relationships". Theor Popul Biol 61(4) :391–408.

[Gribskov *et al.*, 1984] Gribskov, M., Devereux, J., and Burgess, R. R. (1984). "The codon preference plot : graphic analysis of protein coding sequences and prediction of gene expression". Nucleic Acids Res 12(1 Pt 2) :539–49.

[Griffiths-Jones *et al.*, 2003] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). "Rfam : an RNA family database". Nucleic Acids Res 31(1) :439–41.

[Guibourdenche *et al.*, 1986] Guibourdenche, M., Popoff, M. Y., and Riou, J. Y. (1986). "Deoxyribonucleic acid relatedness among Neisseria gonorrhoeae, N. meningitidis, N. lactamica, N. cinerea and "Neisseria polysaccharea"". Ann Inst Pasteur Microbiol 137B(2) :177–85.

[Guigo, 1999] Guigo, R. (1999). "DNA Composition, Codon Usage and Exon Prediction.". In Bishop, M. J., editor, *Genetic Databases*, chapter 8. Academic Press.

[Guindon & Perrière, 2001] Guindon, S., and Perrière, G. (2001). "Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes". Mol Biol Evol 18(9) :1838–40.

[Guénoche, 2003] Guénoche, A. (2003). "Partitions optimisées selon différentes critères :évaluation et comparaison.". Math. Sci. hum., Mathematics and Social Sciences 161 :41–68.

[Guénoche, 2004] Guénoche, A. (2004). "Classification par densité.". Math. Sci. hum., Mathematics and Social Sciences submitted.

[Guénoche & Lescot, 2002] Guénoche, A., and Lescot, M. (2002). "Extraction de classes : application à des données d'expression de gènes.". In *Journées Ouvertes Biologie Informatique Mathématiques Saint-Malo France*, pages 357–364.

[Gupta & Griffiths, 2002] Gupta, R. S., and Griffiths, E. (2002). "Critical issues in bacterial phylogeny". Theor Popul Biol 61(4) :423–34.

[Haas *et al.*, 2001] Haas, L. M. *et al.* (2001). "DiscoveryLink : A system for integrated access to life sciences data sources - Author bios.". IBM Syst J 40 :889–511.

[Hacker & Carniel, 2001] Hacker, J., and Carniel, E. (2001). "Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes". EMBO Rep 2(5) :376–81.

[Hacker & Kaper, 2000] Hacker, J., and Kaper, J. B. (2000). "Pathogenicity islands and the evolution of microbes". Annu Rev Microbiol 54 :641–79.

[Haiech, 2002] Haiech, J. (2002). "Une histoire de la bio-informatique génomique en France.". Médecine / Sciences 18 :131–33.

[Hamoen *et al.*, 1995] Hamoen, L. W., Eshuis, H., Jongbloed, J., Venema, G., and van Sinderen, D. (1995). "A small gene, designated comS, located within the coding region of the fourth amino acid-activation domain of srfA, is required for competence development in Bacillus subtilis". Mol Microbiol 15(1) :55–63.

[Harris, 1997] Harris, N. L. (1997). "Genotator : a workbench for sequence annotation". Genome Res 7(7) :754–62.

[Hartigan & Wong, 1979] Hartigan, J. A., and Wong, M. A. (1979). "A K-means clustering algorithm.". Applied Statistics 28 :100–108.

[Hayes & Borodovsky, 1998a] Hayes, W. S., and Borodovsky, M. (1998a). "Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction". Pac Symp Biocomput pages 279–90.

[Hayes & Borodovsky, 1998b] Hayes, W. S., and Borodovsky, M. (1998b). "How to interpret an anonymous bacterial genome : machine learning approach to gene identification". Genome Res 8(11) :1154–71.

[Hertz & Stormo, 1999] Hertz, G. Z., and Stormo, G. D. (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences". Bioinformatics 15(7-8) :563–77.

[Hertz-Fowler *et al.*, 2004] Hertz-Fowler, C. *et al.* (2004). "GeneDB : a resource for prokaryotic and eukaryotic organisms". Nucleic Acids Res 32 Database issue :D339–43.

[Holm, 1986] Holm, L. (1986). "Codon usage and gene expression". Nucleic Acids Res 14(7) :3075–87.

[Hood, 1999] Hood, D. W. (1999). "The utility of complete genome sequences in the study of pathogenic bacteria". Parasitology 118 Suppl :S3–9.

[Hsiao *et al.*, 2003] Hsiao, W., Wan, I., Jones, S. J., and Brinkman, F. S. (2003). "IslandPath : aiding detection of genomic islands in prokaryotes". Bioinformatics 19(3) :418–20.

[Hubbard *et al.*, 2002] Hubbard, T. *et al.* (2002). "The Ensembl genome database project". Nucleic Acids Res 30(1) :38–41.

[Hutchinson, 1996] Hutchinson, F. (1996). "Mutagenesis.". In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 118. American Society for Microbiology Press.

[Ikemura, 1985] Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms". Mol Biol Evol 2(1) :13–34.

[Iliopoulos *et al.*, 2003] Iliopoulos, I. *et al.* (2003). "Evaluation of annotation strategies using an entire genome sequence". Bioinformatics 19(6) :717–26.

[Iliopoulos *et al.*, 2001] Iliopoulos, I. *et al.* (2001). "Genome sequences and great expectations". Genome Biol 2(1) :INTERACTIONS0001.

[IUBMB, 1992] IUBMB (1992). “Enzyme Nomenclature : Recommendations (1992) of the Nomenclature Committee of the Internationnal Union of Biochemistry and Molecular Biology.”. Academic Press, San Diego, CA.

[Jackson, 2002] Jackson, A. (2002). “Intégration d’une méthode de ré-annotation de génomes bactériens dans une base de données : Application aux génomes des *Salmonella*.”. Master business engineering bio-informatique, Ecole de Biologie Industrielle.

[Jain *et al.*, 2002] Jain, R., Rivera, M. C., Moore, J. E., and Lake, J. A. (2002). “Horizontal gene transfer in microbial genome evolution”. Theor Popul Biol 61(4) :489–95.

[Kanehisa, 2002] Kanehisa, M. (2002). “The KEGG database”. Novartis Found Symp 247 :91–101 ; discussion 101–3, 119–28, 244–52.

[Kanehisa *et al.*, 1984] Kanehisa, M., Fickett, J. W., and Goad, W. B. (1984). “A relational database system for the maintenance and verification of the Los Alamos sequence library”. Nucleic Acids Res 12(1 Pt 1) :149–58.

[Kanehisa *et al.*, 2004] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). “The KEGG resource for deciphering the genome”. Nucleic Acids Res 32 Database issue :D277–80.

[Karlin, 2001] Karlin, S. (2001). “Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes”. Trends Microbiol 9(7) :335–43.

[Karlin *et al.*, 2002] Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B. E., and Mrazek, J. (2002). “Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes”. Theor Popul Biol 61(4) :367–90.

[Karlin *et al.*, 2001] Karlin, S., Mrazek, J., Campbell, A., and Kaiser, D. (2001). “Characterizations of highly expressed genes of four fast-growing bacteria”. J Bacteriol 183(17) :5025–40.

[Karlin *et al.*, 1998] Karlin, S., Mrazek, J., and Campbell, A. M. (1998). “Codon usages in different gene classes of the Escherichia coli genome”. Mol Microbiol 29(6) :1341–55.

[Karp *et al.*, 2002] Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). “The MetaCyc Database”. Nucleic Acids Res 30(1) :59–61.

[Keener & Nomura, 1996] Keener, J., and Nomura, M. (1996). “Regulation of Ribosome Synthesis.”. In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 90. American Society for Microbiology Press.

[Kimura, 1983] Kimura, M. (1983). “The neutral theory of molecular evolution.”. Cambridge University Press, Cambridge.

[Kobayashi *et al.*, 2003] Kobayashi, K. *et al.* (2003). “Essential Bacillus subtilis genes”. Proc Natl Acad Sci U S A 100(8) :4678–83.

[Koonin, 2002] Koonin, E. (2002). “Sequence _ Evolution _ Function”. KLUWER ACADEMIC PUBLISHERS.

[Korbel *et al.*, 2002] Korbel, J. O., Snel, B., Huynen, M. A., and Bork, P. (2002). "SHOT : a web server for the construction of genome phylogenies". Trends Genet 18(3) :158–62.

[Koski *et al.*, 2001] Koski, L. B., Morton, R. A., and Golding, G. B. (2001). "Codon bias and base composition are poor indicators of horizontally transferred genes". Mol Biol Evol 18(3) :404–12.

[Krieger *et al.*, 2004] Krieger, C. J. *et al.* (2004). "MetaCyc : a multiorganism database of metabolic pathways and enzymes". Nucleic Acids Res 32 Database issue :D438–42.

[Krogh *et al.*, 2001] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). "Predicting transmembrane protein topology with a hidden Markov model : application to complete genomes". J Mol Biol 305(3) :567–80.

[Krogh *et al.*, 1994] Krogh, A., Mian, I. S., and Haussler, D. (1994). "A hidden Markov model that finds genes in E. coli DNA". Nucleic Acids Res 22(22) :4768–78.

[Kulikova *et al.*, 2004] Kulikova, T. *et al.* (2004). "The EMBL Nucleotide Sequence Database". Nucleic Acids Res 32 Database issue :D27–30.

[Kumar *et al.*, 2001] Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001). "MEGA2 : molecular evolutionary genetics analysis software". Bioinformatics 17(12) :1244–5.

[Kunisawa *et al.*, 1998] Kunisawa, T., Kanaya, S., and Kutter, E. (1998). "Comparison of synonymous codon distribution patterns of bacteriophage and host genomes". DNA Res 5(6) :319–26.

[Kunst *et al.*, 1997] Kunst, F. *et al.* (1997). "The complete genome sequence of the gram-positive bacterium Bacillus subtilis". Nature 390(6657) :249–56.

[Kurland *et al.*, 2003] Kurland, C. G., Canback, B., and Berg, O. G. (2003). "Horizontal gene transfer : a critical view". Proc Natl Acad Sci U S A 100(17) :9658–62.

[Labarre, 2000] Labarre, L. (2000). "Interfaçage et mise en ligne de PkGDB : une base de données multigénomes.". Dess informatique appliquée à la biologie, UVSQ Versailles–UPMC Paris VI.

[Labarre *et al.*, 2003] Labarre, L., Bocs, S., Derbise, A., Carniel, E., and Médigue, C. (2003). "Investigating *Yersinia pestis* virulence strategies using comparative genomics.". Proceedings of the European Conference on Prokaryotic Genomes (ECPG'2003), Gottingen (Allemagne).

[Labarre & Médigue, 2004] Labarre, L., and Médigue, C. (2004). "The Syntonizer.". Nucleic Acids Research submitted.

[Lafontaine & Lavery, 2000] Lafontaine, I., and Lavery, R. (2000). "ADAPT : a molecular mechanics approach for studying the structural properties of long DNA sequences". Biopolymers 56(4) :292–310.

[Larsen & Krogh, 2003] Larsen, T. S., and Krogh, A. (2003). "EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance". BMC Bioinformatics 4(1) :21.

[Lawrence, 2003] Lawrence, J. (2003). "When ELFs are ORFs, but don't act like them". Trends Genet 19(3) :131–2.

[Lawrence, 1997] Lawrence, J. G. (1997). “Selfish operons and speciation by gene transfer”. Trends Microbiol 5(9) :355–9.

[Lawrence, 2002] Lawrence, J. G. (2002). “Gene transfer in bacteria : speciation without species ?”. Theor Popul Biol 61(4) :449–60.

[Lawrence & Hendrickson, 2003] Lawrence, J. G., and Hendrickson, H. (2003). “Lateral gene transfer : when will adolescence end ?”. Mol Microbiol 50(3) :739–49.

[Lawrence & Ochman, 1998] Lawrence, J. G., and Ochman, H. (1998). “Molecular archaeology of the Escherichia coli genome”. Proc Natl Acad Sci U S A 95(16) :9413–7.

[Lawrence & Ochman, 2002] Lawrence, J. G., and Ochman, H. (2002). “Reconciling the many faces of lateral gene transfer”. Trends Microbiol 10(1) :1–4.

[Lebart *et al.*, 2000] Lebart, L., Piron, M., and Morineau, A. (2000). “Statistique exploratoire multidimensionnelle.”. Dunod.

[Lefebvre, 1999] Lefebvre, C. (1999). “Développement d’une base de données multigénomes.”. Diplôme informatique appliquée, Université d’Orsay, Paris XI.

[Lewis *et al.*, 2002] Lewis, S. E. *et al.* (2002). “Apollo : a sequence annotation editor”. Genome Biol 3(12) :RESEARCH0082.

[Liang *et al.*, 2002] Liang, P., Labedan, B., and Riley, M. (2002). “Physiological genomics of Escherichia coli protein families”. Physiol Genomics 9(1) :15–26.

[Lowe & Eddy, 1997] Lowe, T. M., and Eddy, S. R. (1997). “tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence”. Nucleic Acids Res 25(5) :955–64.

[Lukashin & Borodovsky, 1998] Lukashin, A. V., and Borodovsky, M. (1998). “GeneMark.hmm : new solutions for gene finding”. Nucleic Acids Res 26(4) :1107–15.

[MacQueen, 1967] MacQueen, J. (1967). “Some methods for classification and analysis of multivariate observations.”. In Le Cam, L. M., and Neyman, J., editors, *Proceedings of the Fifth Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

[Mantri & Williams, 2004] Mantri, Y., and Williams, K. P. (2004). “Islander : a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities”. Nucleic Acids Res 32 Database issue :D55–8.

[Marchler-Bauer *et al.*, 2003] Marchler-Bauer, A. *et al.* (2003). “CDD : a curated Entrez database of conserved domain alignments”. Nucleic Acids Res 31(1) :383–7.

[Marsan & Sagot, 2000] Marsan, L., and Sagot, M. F. (2000). “Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification”. J Comput Biol 7(3-4) :345–62.

[Martin, 1999] Martin, W. (1999). “Mosaic bacterial chromosomes : a challenge en route to a tree of genomes”. Bioessays 21(2) :99–104.

[Mathe *et al.*, 1999] Mathe, C., Peresetsky, A., Dehais, P., Van Montagu, M., and Rouze, P. (1999). "Classification of Arabidopsis thaliana gene sequences : clustering of coding sequences into two groups according to codon usage improves gene prediction". J Mol Biol 285(5) :1977–91.

[Mayhew & F.-U., 1996] Mayhew, M., and F.-U., H. (1996). "Molecular Chaperone Proteins.". In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 61. American Society for Microbiology Press.

[McClelland *et al.*, 2001] McClelland, M. *et al.* (2001). "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2". Nature 413(6858) :852–6.

[McGee *et al.*, 1999] McGee, D. J., May, C. A., Garner, R. M., Himpsl, J. M., and Mobley, H. L. (1999). "Isolation of Helicobacter pylori genes that modulate urease activity". J Bacteriol 181(8) :2477–84.

[McInerney, 1998] McInerney, J. O. (1998). "Replicational and transcriptional selection on codon usage in Borrelia burgdorferi". Proc Natl Acad Sci U S A 95(18) :10698–703.

[Médigue, 2000a] Médigue, C. (2000a). "Annotation des Génomes de Micro-organismes Pathogènes.". Habilitation à diriger les recherches, Université de Versailles-Saint-Quentin.

[Médigue, 2000b] Médigue, C. (2000b). "Recherche de régions codantes dans les génomes bactériens : détection d'erreurs d'annotation et de *frameshifts* authentiques.". Worshop on High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) (Swiss Institute of Bioinformatics (SIB)), Grenoble (France).

[Médigue, 2001] Médigue, C. (2001). "La stratégie AMIGene pour l'annotation syntaxique des séquences génomiques bactériennes.". Worshop HAMAP (SIB), Genève (Suisse).

[Médigue *et al.*, 1990] Médigue, C., Henaut, A., and Danchin, A. (1990). "Escherichia coli molecular genetic map (1000 kbp) : update I". Mol Microbiol 4(9) :1443–54.

[Médigue *et al.*, 1999a] Médigue, C., Rechenmann, F., Danchin, A., and Viari, A. (1999a). "Imagene : an integrated computer environment for sequence annotation and analysis". Bioinformatics 15(1) :2–15.

[Médigue *et al.*, 1999b] Médigue, C., Rose, M., Viari, A., and Danchin, A. (1999b). "Detecting and analyzing DNA sequencing errors : toward a higher quality of the Bacillus subtilis genome sequence". Genome Res 9(11) :1116–27.

[Médigue *et al.*, 1991] Médigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991). "Evidence for horizontal gene transfer in Escherichia coli speciation". J Mol Biol 222(4) :851–6.

[Médigue *et al.*, 2002] Médigue, C., Wong, B. C., Lin, M. C., Bocs, S., and Danchin, A. (2002). "The secE gene of Helicobacter pylori". J Bacteriol 184(10) :2837–40.

[Mewes *et al.*, 2002] Mewes, H. W. *et al.* (2002). "MIPS : a database for genomes and protein sequences". Nucleic Acids Res 30(1) :31–4.

[Miyazaki *et al.*, 2004] Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. (2004). "DDBJ in the stream of various biological data". Nucleic Acids Res 32 Database issue :D31–4.

[Morgat & Rechenmann, 2002] Morgat, A., and Rechenmann, F. (2002). "Modélisation des données biologiques.". Médecine / Sciences 18 :366–374.

[Morton, 1994] Morton, B. R. (1994). "Codon use and the rate of divergence of land plant chloroplast genes". Mol Biol Evol 11(2) :231–8.

[Moszer, 1996] Moszer, I. (1996). "Représentation et Analyse des Génomes : Application au Projet de Séquençage du Génome de Bacillus Subtilis.". Génétique, Université Paris VI - Pierre et Marie Curie.

[Moszer *et al.*, 2002] Moszer, I., Jones, L. M., Moreira, S., Fabry, C., and Danchin, A. (2002). "SubtiList : the reference database for the Bacillus subtilis genome". Nucleic Acids Res 30(1) :62–5.

[Moszer *et al.*, 1999] Moszer, I., Rocha, E. P., and Danchin, A. (1999). "Codon usage and lateral gene transfer in Bacillus subtilis". Curr Opin Microbiol 2(5) :524–8.

[Mount, 2001] Mount, D. W. (2001). "Bioinformatics". COLD SPRING HARBOR LABORATORY PRESS.

[Mulder *et al.*, 2003] Mulder, N. J. *et al.* (2003). "The InterPro Database, 2003 brings increased coverage and new features". Nucleic Acids Res 31(1) :315–8.

[Natale *et al.*, 2000] Natale, D. A., Galperin, M. Y., Tatusov, R. L., and Koonin, E. V. (2000). "Using the COG database to improve gene recognition in complete genomes". Genetica 108(1) :9–17.

[Needleman & Wunsch, 1970] Needleman, S. B., and Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". J Mol Biol 48(3) :443–53.

[Nei & Kumar, 2000a] Nei, M., and Kumar, S. (2000a). "Evolutionary Change of DNA Sequences". In *Molecular Evolution and Phylogenetics*, chapter 3. Oxford University Press.

[Nei & Kumar, 2000b] Nei, M., and Kumar, S. (2000b). "Molecular Evolution and Phylogenetics.". Oxford University Press.

[Nei & Kumar, 2000c] Nei, M., and Kumar, S. (2000c). "Phylogenetic Inference : Distance Methods". In *Molecular Evolution and Phylogenetics*, chapter 6. Oxford University Press.

[Nicolas, 2003] Nicolas, P. (2003). "Mise au point et utilisation des modèles de Markov cachés.". Sdv, Université d'Evry Val d'Essonne.

[Nicolas *et al.*, 2002] Nicolas, P. *et al.* (2002). "Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models". Nucleic Acids Res 30(6) :1418–26.

[Nielsen *et al.*, 1997] Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites". Protein Eng 10(1) :1–6.

[Nierman *et al.*, 2001] Nierman, W. C. *et al.* (2001). "Complete genome sequence of Caulobacter crescentus". Proc Natl Acad Sci U S A 98(7) :4136–41.

[Nitschke *et al.*, 1998] Nitschke, P. *et al.* (1998). “Indigo : a World-Wide-Web review of genomes and gene functions”. FEMS Microbiol Rev 22(4) :207–27.

[Nuel, 2001] Nuel, G. (2001). “Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques.”. Mathématiques, Université d'Evry Val d'Essonne.

[Ochman, 2002] Ochman, H. (2002). “Distinguishing the ORFs from the ELFs : short bacterial genes and the annotation of genomes”. Trends Genet 18(7) :335–7.

[Ochman *et al.*, 2000] Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). “Lateral gene transfer and the nature of bacterial innovation”. Nature 405(6784) :299–304.

[Overbeek *et al.*, 1999] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). “Use of contiguity on the chromosome to predict functional coupling”. In Silico Biol 1(2) :93–108.

[Overbeek *et al.*, 2003] Overbeek, R. *et al.* (2003). “The ERGO genome analysis and discovery system”. Nucleic Acids Res 31(1) :164–71.

[Page *et al.*, 2000] Page, M. *et al.* (2000). “Représentation de connaissances au moyen de classes et d'associations : le système AROM.”. In *Actes du colloque Langages et Modèles à objets (LMO), Mont Saint-Hilaire*, pages 91–106. Canada : Editions Hermes.

[Parkhill *et al.*, 2000] Parkhill, J. *et al.* (2000). “Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491”. Nature 404(6777) :502–6.

[Parkhill *et al.*, 2001a] Parkhill, J. *et al.* (2001a). “Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18”. Nature 413(6858) :848–52.

[Parkhill *et al.*, 2003] Parkhill, J. *et al.* (2003). “Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica”. Nat Genet.

[Parkhill *et al.*, 2001b] Parkhill, J. *et al.* (2001b). “Genome sequence of Yersinia pestis, the causative agent of plague”. Nature 413(6855) :523–7.

[Pavesi *et al.*, 1994] Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., and Ottonello, S. (1994). “Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions”. Nucleic Acids Res 22(7) :1247–56.

[Peden, 1999] Peden, J. F. (1999). “Analysis of Codon Usage.”. Department of genetics, University of Nottingham.

[Pedersen *et al.*, 2000] Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H., and Ussery, D. W. (2000). “A DNA structural atlas for Escherichia coli”. J Mol Biol 299(4) :907–30.

[Pelmont, 1993] Pelmont, J. (1993). “Bactéries et environnement, Adaptation physiologiques.”. Presse Universitaire de Grenoble.

[Perna *et al.*, 2001] Perna, N. T. *et al.* (2001). “Genome sequence of enterohaemorrhagic Escherichia coli O157 :H7”. Nature 409(6819) :529–33.

[Perrière, 2000] Perrière, G. (2000). “Bases de données et outils d’analyse pour la génomique bactérienne.”. Habilitation à diriger les recherches, Université Claude Bernard - Lyon 1.

[Perrière *et al.*, 2000a] Perrière, G., Bessières, P., and Labedan, B. (2000a). “EMGLib : the enhanced microbial genomes library (update 2000)”. Nucleic Acids Res 28(1) :68–71.

[Perrière *et al.*, 2000b] Perrière, G., Duret, L., and Gouy, M. (2000b). “HOBACGEN : database system for comparative genomics in bacteria”. Genome Res 10(3) :379–85.

[Perrière & Thioulouse, 2002] Perrière, G., and Thioulouse, J. (2002). “Use and misuse of correspondence analysis in codon usage studies”. Nucleic Acids Res 30(20) :4548–55.

[Peterson *et al.*, 2001] Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001). “The Comprehensive Microbial Resource”. Nucleic Acids Res 29(1) :123–5.

[Pruitt *et al.*, 2003] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2003). “NCBI Reference Sequence project : update and current status”. Nucleic Acids Res 31(1) :34–7.

[R Development Core Team, 2003] R Development Core Team (2003). “R : A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

[Rabiner, 1989] Rabiner, L. R. (1989). “A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition.”. In *Proceedings of the IEEE*, volume 77, pages 257–286.

[Ragan, 2001a] Ragan, M. A. (2001a). “Detection of lateral gene transfer among microbial genomes”. Curr Opin Genet Dev 11(6) :620–6.

[Ragan, 2001b] Ragan, M. A. (2001b). “On surrogate methods for detecting lateral gene transfer”. FEMS Microbiol Lett 201(2) :187–91.

[Ragan & Charlebois, 2002] Ragan, M. A., and Charlebois, R. L. (2002). “Distributional profiles of homologous open reading frames among bacterial phyla : implications for vertical and lateral transmission”. Int J Syst Evol Microbiol 52(Pt 3) :777–87.

[Rappe & Giovannoni, 2003] Rappe, M. S., and Giovannoni, S. J. (2003). “The uncultured microbial majority”. Annu Rev Microbiol 57 :369–94.

[Read *et al.*, 2003] Read, T. D. *et al.* (2003). “The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria”. Nature 423(6935) :81–6.

[Rechenmann & Uvietta, 1991] Rechenmann, F., and Uvietta, P. (1991). “SHIRKA : an object-centered knowledge based management system.”. In Pavé, A., and Vansteenkiste, G., editors, *Artificial intelligence in numerical and symbolic simulation*, pages 9–23. Lyon, France : Aléas.

[Reese, 2001] Reese, M. G. (2001). “Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome”. Comput Chem 26(1) :51–6.

[Rice *et al.*, 2000] Rice, P., Longden, I., and Bleasby, A. (2000). “EMBOSS : the European Molecular Biology Open Software Suite”. Trends Genet 16(6) :276–7.

[Richard & Nuel, 2003] Richard, H., and Nuel, G. (2003). “SPA : simple web tool to assess statistical significance of DNA patterns”. Nucleic Acids Res 31(13) :3679–81.

[Riley & Labedan, 1997] Riley, M., and Labedan, B. (1997). “Protein evolution viewed through Escherichia coli protein sequences : introducing the notion of a structural segment of homology, the module”. J Mol Biol 268(5) :857–68.

[Rocha, 2000] Rocha, E. P. (2000). “Analyse exploratoire des génomes bactériens.”. Génétique cellulaire et moléculaire, Université de Versailles Saint-Quentin-En-Yvelines.

[Rocha & Danchin, 2003] Rocha, E. P., and Danchin, A. (2003). “Essentiality, not expressiveness, drives gene-strand bias in bacteria”. Nat Genet 34(4) :377–8.

[Rocha *et al.*, 1999] Rocha, E. P., Danchin, A., and Viari, A. (1999). “Translation in Bacillus subtilis : roles and trends of initiation and termination, insights from a genome analysis”. Nucleic Acids Res 27(17) :3567–76.

[Rojas *et al.*, 2003] Rojas, A., Garcia-Vallve, S., Montero, M. A., Arola, L., and Romeu, A. (2003). “Frameshift mutation events in beta-glucosidases”. Gene 314 :191–9.

[Romanet, 2001] Romanet, J. (2001). “Recherche de zones codantes sur un génome complet”. Dea informatiques : Systèmes et communications, Université Grenoble I - Joseph Fourier.

[Roth *et al.*, 1996] Roth, J. R. *et al.* (1996). “Rearrangements of the Bacterial Chromosome : Formation and Applications.”. In Neidhardt, F. C. *et al.*, editors, *Escherichia coli et Salmonella*, chapter 120. American Society for Microbiology Press.

[Rudd, 1998] Rudd, K. E. (1998). “Linkage map of Escherichia coli K-12, edition 10 : the physical map”. Microbiol Mol Biol Rev 62(3) :985–1019.

[Rudd, 2000] Rudd, K. E. (2000). “EcoGene : a genome sequence database for Escherichia coli K-12”. Nucleic Acids Res 28(1) :60–4.

[Rutherford *et al.*, 2000] Rutherford, K. *et al.* (2000). “Artemis : sequence visualization and annotation”. Bioinformatics 16(10) :944–5.

[Salgado *et al.*, 2004] Salgado, H. *et al.* (2004). “RegulonDB (version 4.0) : transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12”. Nucleic Acids Res 32 Database issue :D303–6.

[Salgado *et al.*, 2001] Salgado, H. *et al.* (2001). “RegulonDB (version 3.2) : transcriptional regulation and operon organization in Escherichia coli K-12”. Nucleic Acids Res 29(1) :72–4.

[Salzberg *et al.*, 1998] Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). “Microbial gene identification using interpolated Markov models”. Nucleic Acids Res 26(2) :544–8.

[Sanger *et al.*, 1977] Sanger, F. *et al.* (1977). “Nucliotide sequence of bacteriophage phi X174 DNA”. Nature 265(5596) :687–95.

[Saunders *et al.*, 1999] Saunders, N. J., Hood, D. W., and Moxon, E. R. (1999). “Bacterial evolution : bacteria play pass the gene”. Curr Biol 9(5) :R180–3.

[Saunders *et al.*, 2000] Saunders, N. J. *et al.* (2000). "Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58". Mol Microbiol 37(1) :207–15.

[Schbath, 1997] Schbath, S. (1997). "An efficient statistic to detect over- and under-represented words in DNA sequences". J Comput Biol 4(2) :189–92.

[Schiex *et al.*, 2003] Schiex, T., Gouzy, J., Moisan, A., and de Oliveira, Y. (2003). "FrameD : a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences". Nucleic Acids Res 31(13) :3738–41.

[Schiex *et al.*, 2000] Schiex, T., Thébault, P., and Kahn, D. (2000). "Recherche des gènes et des erreurs de séquençage dans les génomes bactériens GC-riches (et autres...).". In *Proc. of JOBIM'2000 Montpellier France*, pages 321–328.

[Schomburg *et al.*, 2004] Schomburg, I. *et al.* (2004). "BRENDA, the enzyme database : updates and major new developments". Nucleic Acids Res 32 Database issue :D431–3.

[Schuler *et al.*, 1996] Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). "Entrez : molecular biology database and retrieval system". Methods Enzymol 266 :141–62.

[Serres *et al.*, 2004] Serres, M. H., Goswami, S., and Riley, M. (2004). "GenProtEC : an updated and improved analysis of functions of Escherichia coli K-12 proteins". Nucleic Acids Res 32 Database issue :D300–2.

[Sharp *et al.*, 1988] Sharp, P. M. *et al.* (1988). "Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens ; a review of the considerable within-species diversity". Nucleic Acids Res 16(17) :8207–11.

[Sharp & Li, 1987] Sharp, P. M., and Li, W. H. (1987). "The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications". Nucleic Acids Res 15(3) :1281–95.

[Sharp *et al.*, 1986] Sharp, P. M., Tuohy, T. M., and Mosurski, K. R. (1986). "Codon usage in yeast : cluster analysis clearly differentiates highly and lowly expressed genes". Nucleic Acids Res 14(13) :5125–43.

[Shepherd, 1981] Shepherd, J. C. W. (1981). "Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification". Proc. Natl. Acad. Sci. USA 78 :1596–1600.

[Shmatkov *et al.*, 1999] Shmatkov, A. M., Melikyan, A. A., Chernousko, F. L., and Borodovsky, M. (1999). "Finding prokaryotic genes by the 'frame-by-frame' algorithm : targeting gene starts and overlapping genes". Bioinformatics 15(11) :874–86.

[Skovgaard *et al.*, 2001] Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. (2001). "On the total number of genes and their length distribution in complete microbial genomes". Trends Genet 17(8) :425–8.

[Smith *et al.*, 1992] Smith, M. W., Feng, D. F., and Doolittle, R. F. (1992). "Evolution by acquisition : the case for horizontal gene transfers". Trends Biochem Sci 17(12) :489–93.

[Smith & Eyre-Walker, 2001] Smith, N. G., and Eyre-Walker, A. (2001). "Why are translationally sub-optimal synonymous codons used in Escherichia coli ?". J Mol Evol 53(3) :225–36.

[Smith & Waterman, 1981] Smith, T. F., and Waterman, M. S. (1981). "Identification of common molecular subsequences". J Mol Biol 147(1) :195–7.

[Sneath & Sokal, 1973] Sneath, P. H. A., and Sokal, R. R. (1973). "Numerical Taxonomy.". Freeman, San Francisco, CA.

[Sprinzl *et al.*, 1998] Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998). "Compilation of tRNA sequences and sequences of tRNA genes". Nucleic Acids Res 26(1) :148–53.

[Stackebrandt *et al.*, 2002] Stackebrandt, E. *et al.* (2002). "Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology". Int J Syst Evol Microbiol 52(Pt 3) :1043–7.

[Stein, 2001] Stein, L. (2001). "Genome annotation : from sequence to biology". Nat Rev Genet 2(7) :493–503.

[Stevens *et al.*, 2000] Stevens, R. *et al.* (2000). "TAMBIS : transparent access to multiple bioinformatics information sources". Bioinformatics 16(2) :184–5.

[Strick *et al.*, 1998] Strick, T. R., Croquette, V., and Bensimon, D. (1998). "Homologous pairing in stretched supercoiled DNA". Proc Natl Acad Sci U S A 95(18) :10579–83.

[Sueoka, 1993] Sueoka, N. (1993). "Directional mutation pressure, mutator mutations, and dynamics of molecular evolution". J Mol Evol 37(2) :137–53.

[Suzek *et al.*, 2001] Suzek, B. E., Ermolaeva, M. D., Schreiber, M., and Salzberg, S. L. (2001). "A probabilistic method for identifying start codons in bacterial genomes". Bioinformatics 17(12) :1123–30.

[Syvanen, 1985] Syvanen, M. (1985). "Cross-species gene transfer ; implications for a new theory of evolution". J Theor Biol 112(2) :333–43.

[Syvanen, 1994] Syvanen, M. (1994). "Horizontal gene transfer : evidence and possible consequences". Annu Rev Genet 28 :237–61.

[Takami *et al.*, 2001] Takami, H., Han, C. G., Takaki, Y., and Ohtsubo, E. (2001). "Identification and distribution of new insertion sequences in the genome of alkaliphilic Bacillus halodurans C-125". J Bacteriol 183(14) :4345–56.

[Takami & Horikoshi, 2000] Takami, H., and Horikoshi, K. (2000). "Analysis of the genome of an alkaliphilic Bacillus strain from an industrial point of view". Extremophiles 4(2) :99–108.

[Tatusov *et al.*, 2001] Tatusov, R. L. *et al.* (2001). "The COG database : new developments in phylogenetic classification of proteins from complete genomes". Nucleic Acids Res 29(1) :22–8.

[Tettelin *et al.*, 2001] Tettelin, H. *et al.* (2001). "Complete genome sequence of a virulent isolate of Streptococcus pneumoniae". Science 293(5529) :498–506.

[Tettelin *et al.*, 2000] Tettelin, H. *et al.* (2000). "Complete genome sequence of Neisseria meningitidis serogroup B strain MC58". Science 287(5459) :1809–15.

[Uchiyama, 2003] Uchiyama, I. (2003). "MBGD : microbial genome database for comparative analysis". Nucleic Acids Res 31(1) :58–62.

[Vallenet, 2002] Vallenet, D. (2002). "Développement d'une base de données pour l'étude des génomes microbiens pathogènes : PATHODB.". Dess etude des génomes : Outils informatiques et statistiques, Université de Rouen – Faculté des Sciences et Techniques – CFA.

[van Helden, 2003] van Helden, J. (2003). "Regulatory sequence analysis tools". Nucleic Acids Res 31(13) :3593–6.

[Vincent, 2001] Vincent, J. (2001). "Développement d'une méthode d'annotation automatique de phase codantes dans les génomes procaryotes.". Ingénieur informaticien deuxième année, CNAM Institut d'Informatique d'Entreprise.

[Volle, 1997] Volle, M. (1997). "Analyse de données.". Economie et Statistiques avancées. Economica.

[Waldor & RayChaudhuri, 2000] Waldor, M. K., and RayChaudhuri, D. (2000). "Treasure trove for cholera research". Nature 406(6795) :469–70.

[Walker & Koonin, 1997] Walker, D. R., and Koonin, E. V. (1997). "SEALS : a system for easy analysis of lots of sequences". Proc Int Conf Intell Syst Mol Biol 5 :333–9.

[Wassarman & Storz, 2000] Wassarman, K. M., and Storz, G. (2000). "6S RNA regulates E. coli RNA polymerase activity". Cell 101(6) :613–23.

[Wassenaar & Gaastra, 2001] Wassenaar, T. M., and Gaastra, W. (2001). "Bacterial virulence : can we draw the line ?". FEMS Microbiol Lett 201(1) :1–7.

[Wei *et al.*, 2001] Wei, Y. *et al.* (2001). "High-density microarray-mediated gene expression profiling of Escherichia coli". J Bacteriol 183(2) :545–56.

[Woese, 2000] Woese, C. R. (2000). "Interpreting the universal phylogenetic tree". Proc Natl Acad Sci U S A 97(15) :8392–6.

[Wootton & Federhen, 1993] Wootton, J., and Federhen, S. (1993). "Statistics of local complexity in amino acid sequences and sequence databases.". Comput. Chem. 17 :149–163.

[Wren, 2000] Wren, B. W. (2000). "Microbial genome analysis : insights into virulence, host adaptation and evolution". Nat Rev Genet 1(1) :30–9.

[Wright, 1990] Wright, F. (1990). "The 'effective number of codons' used in a gene". Gene 87(1) :23–9.

[Wu *et al.*, 2002] Wu, C. H. *et al.* (2002). "The Protein Information Resource : an integrated public resource of functional annotation of proteins". Nucleic Acids Res 30(1) :35–7.

[Wu & Manber, 1991] Wu, S., and Manber, U. (1991). "Fast Text Searching With Errors.". Technical Report TR-91-11, University of Arizona, Department of Computer Science.

[Yada *et al.*, 1998] Yada, T., Totoki, Y., Ishikawa, M., Asai, K., and Nakai, K. (1998). "Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences". Bioinformatics 14(4) :317–25.

[Zdobnov & Apweiler, 2001] Zdobnov, E. M., and Apweiler, R. (2001). "InterProScan–an integration platform for the signature-recognition methods in InterPro". Bioinformatics 17(9) :847–8.

[Zwieb *et al.*, 1998] Zwieb, C., Larsen, N., and Wower, J. (1998). "The tmRNA database (tmRDB)". Nucleic Acids Res 26(1) :166–7.

Annexes

Annexe A

Biologie des génomes procaryotes

A.1 Reconstruction d'arbre phylogénétique

A.1.1 ARNr 16S

Les séquences d'ARNr 16S ont été extraites à partir de fichiers d'annotation des génomes complets de procaryotes disponibles dans les banques publiques INSD. Comme extragroupe eucaryote, nous avons choisi l'ARNr 18S de Saccharomyces cerevisiae. Nous avons étudiés treize séquences d'archae et quatre-vingt six séquences de bactéries. Pour les huit génomes qui posaient problème (*e.g.* pas d'annotation d'ARNr dans le fichier d'annotation), nous avons donc recherché la présence d'ARNr 16S sur le chromosome par similarité de séquence avec la méthode FindrRNA (voir p. 77). Les 100 séquences d'ARN 16S ont été alignées avec Clustalw (paramètres par défaut). Puis les arbres phylogénétiques ont été calculés dans le logiciel MEGA2 [Kumar *et al.*, 2001]. La méthode de reconstruction d'arbre est celle du « Neighbor Joining (NJ) » (voir p. 136). Le modèle de distance est celui de Kimura à deux paramètres pour les nucléotides (distances d'alignement deux à deux). Enfin, nous testons la robustesse de l'arbre phylogénétique inféré par un « bootstrap » de 500 réplications. Dans ces conditions deux arbres sont produits par MEGA2 : le meilleur arbre (avec distances) et l'arbre consensus (sans distance). Le meilleur arbre montre toutes les valeurs de « bootstrap », dans notre cas certaines sont très faibles ce qui signifie que certains groupes sont inclassables par l'approche utilisée ici. L'arbre consensus ne classe que les groupes qui ont un bootstrap supérieur à 50. C'est celui qui est présenté ici, la longueur des branches n'est pas corrélée à une distance phylogénétique ou temps de divergence entre les espèces.

A.1.2 Protéomes complets

Les méthodes de reconstruction d'arbre à partir des génomes complets sont généralement fondées sur la recherche de polypeptides (gènes) orthologues entre les protéomes (génomes) : c'est le cas par exemple des arbres reconstruits à partir du contenu en gène (Fig. A.1 p. 384) ou à partir de l'ordre des gènes (Fig. A.2 p. 385).

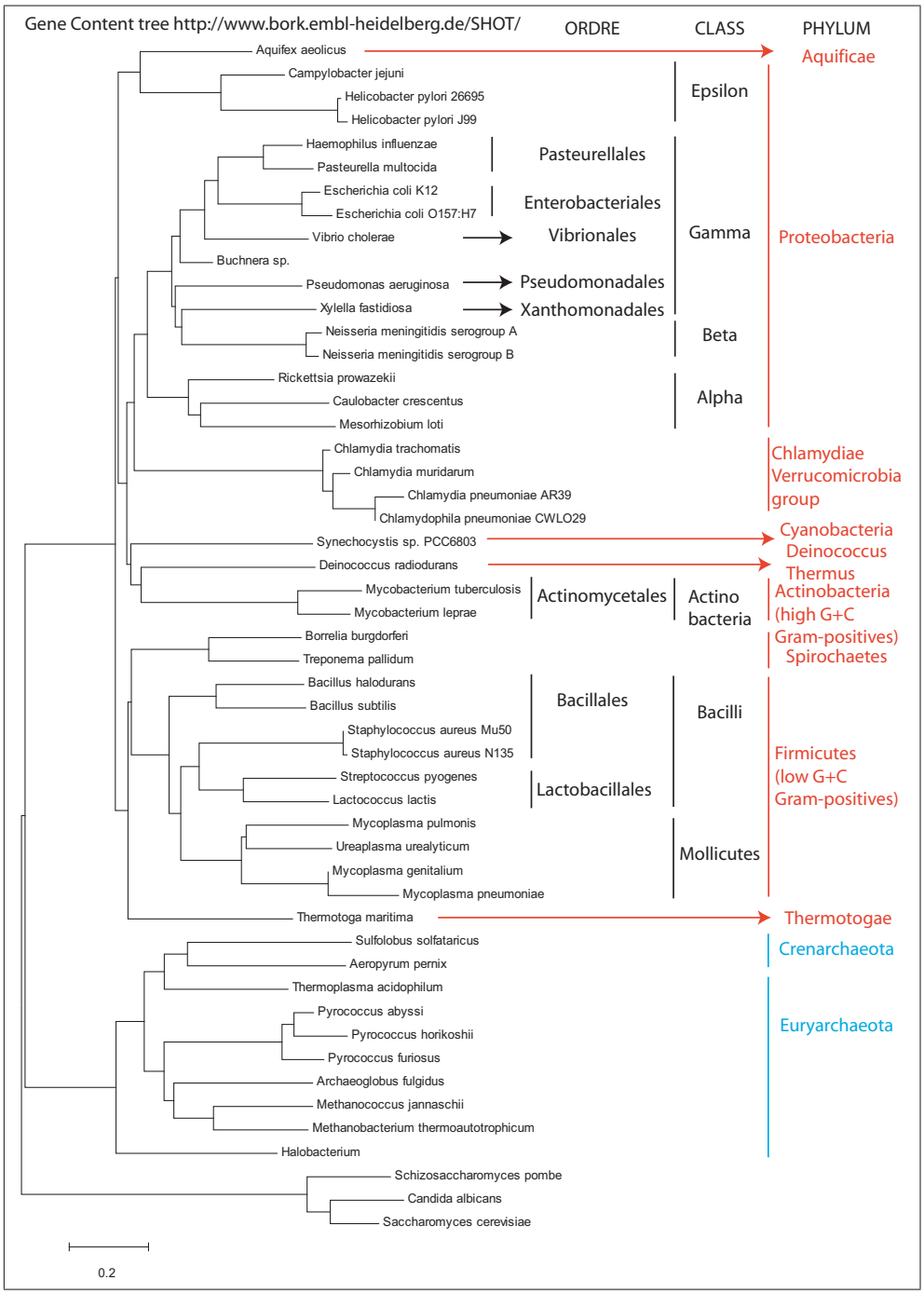


FIG. A.1 – Arbre phylogénomique du contenu en gènes de génomes procaryotes complets

Sur le site de SHOT [Korbel *et al.*, 2002], on choisit la méthode de reconstruction phylogénomique du contenu en gènes et les espèces à classer. Dans notre cas nous avons pris les dix espèces d'archaea et les trente-huit espèces de bactéries ainsi que trois espèces de levures. Les valeurs des autres paramètres sont celles par défaut. Après avoir lancé la procédure on récupère le fichier de la matrice de d'indices de distance. L'indice de distance est calculée par $d = -ln(s)$, s étant la similitude entre deux génomes. La similitude entre deux génomes correspond au nombre de gènes orthologues divisé par le nombre total de gènes du plus petit génomes (qui correspond au nombre maximum de gènes orthologues entre les deux génomes). La matrice d'indices de distance est chargée dans MEGA2 (FIG. 1.2 p. 53). A partir de matrice de distances entre chaque couple, la procédure du NJ construit le meilleur arbre avec les distances phylogénomiques. Comme il n'y a pas de données d'alignement, la robustesse de l'arbre ne peut être testée.

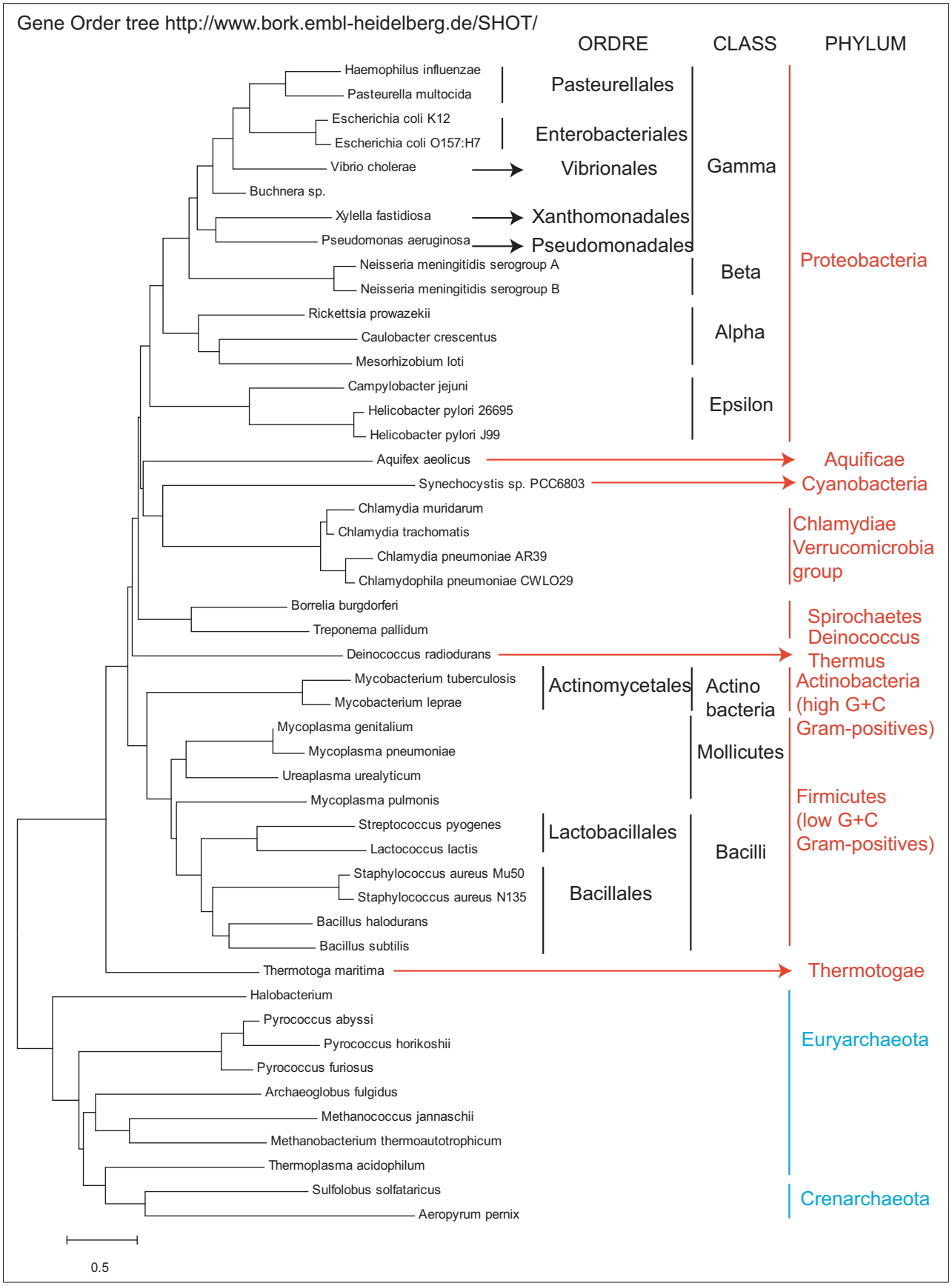


FIG. A.2 – Arbre phylogénomique de l'ordre des gènes de génomes procaryotes complets

C'est le même principe que pour l'arbre de la figure 1.2 p. 53 sauf qu'on choisit la méthode de reconstruction phylogénomique de l'ordre des gènes (analyse de la présence de paires de gènes orthologues co-localisées sur deux chromosomes). Le critère d'ordre des gènes est donc ajouté au critère du contenu en gènes (nombre d'orthologues entre deux génomes).

A.2 Taxonomie et nomenclature bactérienne

Rangs hiérarchiques et nomenclatures types

Rang	Equivalent latin, abréviation	Suffixe	Nomenclature type	Exemple (Pseudomonas syringae)
Domaine ou empire*	*Regio* ou *Imperium*	*-a* **		"*Bacteria*" ou "*Eubacteria*"
Règne***	*Regnum*			"*Bacteria*"
Division ou phylum***	*Divisio* ou *phylum*			*Gracilicutes*
Classe	*Classis* : class.	*-ia* ****	Ordre type	Selon les auteurs : *Proteobacteria* ou *Schizomycetes*
Sous-classe	*Subclassis* : subclass.	*-idae* ****	Ordre type	"Gamma" (si un auteur estime que le genre *Pseudomonas* appartient à la classe des *Proteobacteria*)
Ordre	*Ordo* : ord.	*-ales*	Genre type	*Pseudomonadales*
Sous-ordre	*Subordo* : subord.	*-ineae*	Genre type	*Pseudomonadineae*
Famille	*Familia* : fam.	*-aceae*	Genre type	*Pseudomonadaceae*
Sous-famille	*Subfamilia* : subfam.	*-oideae*	Genre type	"*Pseudomonadoideae******"
Tribu	*Tribus*	*-eae*	Genre type	*Pseudomonadeae*
Sous-tribu	*Subtribus*	*-inae*	Genre type	"*Pseudomonadinae******"
Genre	*Genus* : gen.		Espèce type	*Pseudomonas*
Sous-genre	*Subgenus* : subgen.		Espèce type	Sous-genre "*Pseudomonas******"
Espèce	*Species* : sp.		Souche type	*Pseudomonas syringae*
Sous-espèce	*Subspecies* : subsp.		Souche type	*Pseudomonas syringae* subsp. *syringae*
Biovar***, sérovar***, pathovar***…				pv. Tomato

* : Rang hiérarchique proposé après la dernière édition du code de nomenclature.
** : Suffixe proposé par Woese et al., 1990.
*** : Rangs hiérarchiques non régis par les règles du code de nomenclature.
**** : Suffixes proposés par Stackebrandt et al. (1997) et dont l'utilisation n'est pas obligatoire et ne peut s'appliquer aux classes et sous-classes validement publiées avant 1997.
***** : Exemple théorique car cette nomenclature n'a jamais été proposée.
Tableau extrait de http://www.bacterio.cict.fr/bacdico/systematique/tnomenclature.html

Tab. A.1 – Rangs hiérarchiques et nomenclatures types

Annexe B

Type d'information des fichiers INSD

ID	identification (begins each entry ; 1 per entry)
AC	accession number (>=1 per entry)
SV	sequence version (1 per entry)
DT	date (2 per entry)
DE	description (>=1 per entry)
KW	keyword (>=1 per entry)
OS	organism species (>=1 per entry)
OC	organism classification (>=1 per entry)
OG	organelle (0 or 1 per entry)
RN	reference number (>=1 per entry)
RC	reference comment (>=0 per entry)
RP	reference positions (>=1 per entry)
RX	reference cross-reference (>=0 per entry)
RA	reference author(s) (>=1 per entry)
RT	reference title (>=1 per entry)
RL	reference location (>=1 per entry)
DR	database cross-reference (>=0 per entry)
CC	comments or notes (>=0 per entry)
AH	assembly header (0 or 1 per entry)
AS	assembly information (0 or >=1 per entry)
FH	feature table header (0 or 2 per entry)
FT	feature table data (>=0 per entry)
XX	spacer line (many per entry)
SQ	sequence header (1 per entry)
CO	contig/construct line (0 or >=1 per entry)
bb	(blanks) sequence data (>=1 per entry)
//	termination line (ends each entry ; 1 per entry)

Tab. B.1 – Code à deux lettres indiquant le type d'information contenu dans la ligne de l'entrée de séquence EMBL

Annexe C

Chaînes de Markov pour la prédiction de gènes procaryotes

C.1 *GeneMark*

Cette section résume les travaux originaux de M. Borodovsky *et coll.* sur *GeneMark*, un programme de prédiction de gènes procaryotes fondé sur des modèles de chaînes de Markov [Borodovsky & McIninch, 1993b, Borodovsky & McIninch, 1993a]. Les notations mathématiques ont été modifiées de manière à respecter la cohérence de la thèse et de manière à comprendre les fondements des méthodes de prédiction de gènes par chaînes de Markov. Soit la séquence d'ADN $X_1^n = X_1 \dots X_i \dots X_n,\ X_i \in \mathcal{A}$, où n est la longueur de la séquence. Chaque nucléotide de la séquence est à valeur dans l'alphabet à quatre états $\mathcal{A} = \{A, C, G, T\}$. La principale amélioration entre les deux articles est la suivante :

- dans le premier article, chaque brin est analysé indépendamment (deux modèles sont utilisés π_{non} et π_{pos} [Borodovsky & McIninch, 1993b]) ;
- dans le second article, les deux brins sont analysés simultanément (trois modèles sont utilisés π_{non}, π_{pos} et ω_{pos} [Borodovsky & McIninch, 1993b]).

C.1.1 La première étape

La phase d'apprentissage consiste à estimer les probabilités des matrices de transition. Les régions non-codantes sont modélisées par des chaînes de Markov homogènes. Prenons l'exemple d'un modèle $M1$, les paramètres du modèle sont les probabilités initiales et les probabilités de transition, ajustées sur la statistique des mots de longueur 2. Les valeurs numériques des paramètres (vecteur ou matrice ligne des probabilités initiales μ_{non} et matrice de transition π_{non}) sont estimées à partir des comptages des mono- et dinucléotides N_a et N_{ab} du jeu de séquences d'ADN non-codantes.

Selon le principe de maximum de vraisemblance, on a :

$$\begin{aligned} \hat{\pi}_{non}(a,b) &= \frac{N_{ab}}{N_a} \ \forall a,b \in \mathcal{A},\ pos = 1,2,3 \\ \hat{\mu}_{non}(a) &= \frac{N_a}{n} \end{aligned}$$

. Les probabilités initiales sont donc estimées par les fréquences normalisées en mononucléotides. Nous obtenons donc un modèle qui prend en compte les corrélations entre les nucléotides adjacents, uniformément distribués le long des séquences non-codantes, bien que les patrons peuvent avoir une distribution bien plus compliquée.

Les chaînes de Markov homogènes ne décrivent pas bien les régions codantes d'un point de vue statistique. En revanche les chaînes de Markov non homogènes ou à transition 3-périodique décrivent plus précisément la faible périodicité de trois nucléotides des CDS. Les régions codantes sont donc modélisées par des chaînes de Markov à transition 3-périodique. Les paramètres d'un modèle $M1\_3$ sont trois vecteurs de probabilités initiales $\mu_1,\ \mu_2$ *et* μ_3 et trois matrices de transition $\pi_1,\ \pi_2$ *et* π_3. La valeur de ces paramètres est estimée par maximum de vraisemblance. Le jeu d'apprentissage des CDS est concaténé en une longue séquence de longueur N en excluant les codons stop. En revanche le jeu d'apprentissage des séquences non-codantes n'a pas été concaténé. Les comptages des mono- et dinucléotides sont séparés en trois sous-jeux en fonction de la position $pos = 1, 2, 3$ du nucléotide a dans le codon :

$$\begin{aligned} \hat{\pi}_{pos}(a,b) &= \frac{N_{ab,pos}}{N_{a,pos-1}} \ \forall a,b \in \mathcal{A},\ pos = 1,2,3 \\ \hat{\mu}_{pos}(a) &= \frac{N_{a,pos}}{N/3}. \end{aligned}$$

Une procédure similaire, comptage et séparation en trois sous-jeux pour le m-uple et $m+1$-uple, est accomplie pour estimer la valeur des paramètres de modèles Mm_3. Plus le jeu de données est grand, et plus le modèle sera réaliste car nous pourrons établir la matrice de transition en utilisant un ordre de Markov élevé c'est-à-dire tenant compte d'un contexte important. En fait, le choix de l'ordre du modèle résulte d'un subtil compromis, entre un ordre faible qui permettra une meilleure estimation des paramètres du modèle et un ordre élevé qui rendra le modèle plus réaliste (*e.g.* ne pas dépasser un ordre 5, *i.e.* comptage des dicodons). Cette approche peut être étendue au cas où des nucléotides ambigus à valeur dans les symbols standards : Y,R,M,K,S,W,H,B,D,V,D,N sont présents dans la séquence d'ADN.

Les ombres des régions codantes sont elles aussi facilement modélisables par des chaînes de Markov non-homogènes ou à transition tri-périodique. Réservons la lettre O pour tous les paramètres qui apparaissent dans le modèle $M1\_3$ de l'ombre d'une région codante : probabilités initiales (ν_1, ν_2, ν_3) et probabilités de transition (ω_1, ω_2, ω_3) etc. On peut soit utiliser le jeu d'apprentissage des ombres des « vraies » régions codantes soit trouver $\omega_{pos}(a,b)$ analytiquement en combinant les valeurs des statistiques $N_{ab,pos}$ et $N_{a,pos-1}$ qui sont connues pour l'analyse des « vraies » CDS du jeu d'apprentissage.

Ainsi, nous avons trois modèles, un pour le non-codant, un pour le codant et un pour les ombres du codant, donc en théorie trois jeux d'apprentissage, le jeu des CDS, le jeu des ombres des CDS et le jeu des séquences non-codantes. En pratique la matrice de transitions des ombres de CDS se déduit de celle des CDS. Il est aussi possible de calculer la matrice des séquences non-codantes à partir du jeu de CDS, en faisant la moyenne des probabilités des m-uples des CDS sur les 6 phases, *shuffle*, mais il est préférable d'avoir un jeu représentatif de séquences non-codantes natives. Pour le jeu de CDS et pour le jeu des ombres, lors des comptages il y a séparation en trois sous-jeux (mot de longueur $m+1$ en position 1 des codons, en position 2 ou en position 3).

C.1.2 La seconde étape

La procédure de reconnaissance de CDS dans une nouvelle séquence continue d'ADN est interprétée comme la recherche d'une sous-séquence qui dans une certaine mesure est similaire à une des séquences d'ADN modèles possibles produites par le générateur défini par un modèle de Markov non homogène. La nouvelle séquence d'ADN est analysée par fragment en utilisant une fenêtre glissante dont la longueur est généralement inférieure à la longueur moyenne d'une CDS naturelle. Le problème initial de recherche de sous-séquences codantes se substitue à un jeu de problèmes de reconnaissance de fragments d'ADN qui apparaissent dans la fenêtre glissante. Une fenêtre glissante lit le fragment $X_1^w = X_1 \dots X_i \dots X_w,\ X_i \in \mathcal{A}$ de longueur w (multiple de trois) et permet d'effectuer les calculs de probabilités conditionnelles. Les formules suivantes donnent une idée générale de l'algorithme qui est utilisé dans des modèles de chaînes de Markov d'ordre 1. La valeur de la probabilité que X_1^w apparaisse dans une région non-codante est :

$$\mathbb{P}_{\pi_{non}}(X_1^w \mid NON) = \mu_{non}(x_1)\pi_{non}(x_1,x_2)\pi_{non}(x_2,x_3)\pi_{non}(x_3,x_4)\dots\pi_{non}(x_{w-1},x_w)$$

L'apparition de X_1^w dans une région codante peut être divisé en trois sous-événements mutuellement exclusifs en fonction de la phase du fragment définie par la position dans le codon du premier nucléotide. Les probabilités de ces sous-événements dans le modèle $M1\_3$ est :

$$\begin{aligned}
\mathbb{P}_{\pi}(X_1^w \mid COD_1) &= \mu_1(x_1)\pi_2(x_1,x_2)\pi_3(x_2,x_3)\pi_1(x_3,x_4)\dots\pi_3(x_{w-1},x_w)\\
\mathbb{P}_{\pi}(X_1^w \mid COD_2) &= \mu_2(x_1)\pi_3(x_1,x_2)\pi_1(x_2,x_3)\pi_2(x_3,x_4)\dots\pi_1(x_{w-1},x_w)\\
\mathbb{P}_{\pi}(X_1^w \mid COD_3) &= \mu_3(x_1)\pi_1(x_1,x_2)\pi_2(x_2,x_3)\pi_3(x_3,x_4)\dots\pi_2(x_{w-1},x_w)
\end{aligned}$$

La somme totale des trois composants $\mathbb{P}_{\pi}(X_1^w \mid COD_1)$, $\mathbb{P}_{\pi}(X_1^w \mid COD_2)$ et $\mathbb{P}_{\pi}(X_1^w \mid COD_3)$ donne la valeur de $\mathbb{P}_{\pi}(X_1^w \mid COD)$ la probabilité que le X_1^w apparaisse dans une région codante indépendamment de la position relative de son premier nucléotide à l'intérieur d'un codon.

Il y a trois calculs de probabilités supplémentaires pour tenir compte du fait que le fragment puisse tomber dans l'ombre d'une « vraie » région codante.

$$\begin{aligned}
\mathbb{O}_{\omega}(X_1^w \mid COD_1) &= \nu_1(x_1)\omega_2(x_1,x_2)\omega_3(x_2,x_3)\omega_1(x_3,x_4)\dots\omega_3(x_{w-1},x_w)\\
\mathbb{O}_{\omega}(X_1^w \mid COD_2) &= \nu_2(x_1)\omega_3(x_1,x_2)\omega_1(x_2,x_3)\omega_2(x_3,x_4)\dots\omega_1(x_{w-1},x_w)\\
\mathbb{O}_{\omega}(X_1^w \mid COD_3) &= \nu_3(x_1)\omega_1(x_1,x_2)\omega_2(x_2,x_3)\omega_3(x_3,x_4)\dots\omega_2(x_{w-1},x_w)
\end{aligned}$$

Nous cherchons finalement à déterminer les probabilités *a posteriori* $\mathbb{P}_\pi(COD_f \mid X_1^w)$ et $\mathbb{O}_\omega(COD_f \mid X_1^w)$ qui caractérisent la propriété de codage dans du fragment X_1^w étant lu dans les six phases possibles. $\mathbb{P}(COD \mid X_1^w)$ est la probabilité d'être dans une région codante connaissant le fragment d'ADN lu dans l'expérience. Trois composants $\mathbb{P}(COD_f \mid X_1^w)$ sont déterminés selon la formule de Bayes $f = 1, 2, 3$:

$$\begin{aligned}
\mathbb{P}(COD_f \mid X_1^w) &= \frac{\mathbb{P}(X_1^w \mid COD_f) * \mathbb{P}(COD_f)}{\mathbb{P}(X_1^w)} \\
\mathbb{P}(X_1^w) &= \sum_Y \mathbb{P}(X_1^w, Y) = \sum_Y \mathbb{P}(X_1^w \mid Y) * \mathbb{P}(Y) \\
&= \sum_{j=1,2,3} \mathbb{P}_\pi(X_1^w \mid COD_j) * \mathbb{P}(COD_j) \; + \sum_{j=1,2,3} \mathbb{O}_\omega(X_1^w \mid COD_j) * \mathbb{O}(COD_j) \\
&\quad + \mathbb{P}_{\pi_{non}}(X_1^w \mid NON) * \mathbb{P}(NON) \\
\mathbb{P}(COD_f \mid X_1^w) &= \frac{\mathbb{P}_\pi(X_1^w \mid COD_f) * \mathbb{P}(COD_f)}{\sum_{j=1,2,3} \mathbb{P}_\pi(X_1^w \mid COD_j) * \mathbb{P}(COD_j) \; + \; \sum_{j=1,2,3} \mathbb{O}_\omega(X_1^w \mid COD_j) * \mathbb{O}(COD_j)} \\
&\quad \overline{+ \mathbb{P}_{\pi_{non}}(X_1^w \mid NON) * \mathbb{P}(NON)}.
\end{aligned}$$

$\mathbb{P}(COD_f)$ est la probabilité *a priori* de l'événement $COD_f,\ f = 1, 2, 3$ que tout fragment non encore analysé tombe dans une région codante (et que son premier nucléotide soit situé à une position dans le codon définie par l'index f). $\mathbb{O}(COD_f)$ est la probabilité *a priori* de l'événement $COD_f,\ f = 1, 2, 3$ que tout fragment non encore analysé tombe dans une ombre de région codante (et que son premier nucléotide soit situé à une position dans le codon définie par l'index f). $\mathbb{P}(NON)$ est la probabilité *a priori* de l'événement NON, c'est-à-dire que tout fragment non encore analysé tombe dans une région non-codante. L'hypothèse naturelle est que $\mathbb{P}(NON) = 1/2$ et que $\mathbb{P}(COD_f) = \mathbb{O}(COD_f) = 1/12\ f = 1, 2, 3$.

Le même genre de formule définit les différentes phases d'une ombre quand le fragment F est observé :

$$\begin{aligned}
\mathbb{O}(COD_f \mid X_1^w) &= \frac{\mathbb{O}_\omega(X_1^w \mid COD_f) * \mathbb{O}(COD_f)}{\sum_{j=1,2,3} \mathbb{P}_\pi(X_1^w \mid COD_j) * \mathbb{P}(COD_j) \; + \; \sum_{j=1,2,3} \mathbb{O}(X_1^w \mid COD_j) * \mathbb{O}(COD_j)} \\
&\quad \overline{+ \mathbb{P}_{\pi_{non}}(X_1^w \mid NON) * \mathbb{P}(NON)}.
\end{aligned}$$

Les deux dernières équations déterminent les six probabilités *a posteriori* de codage en phase pour n'importe quel fragment X_1^w donné de la nouvelle séquence d'ADN.

La valeur :

$$\begin{aligned}
\mathbb{P}(NON \mid X_1^w) &= \frac{\mathbb{P}_{\pi_{non}}(X_1^w \mid NON) * \mathbb{P}(NON)}{\sum_{j=1,2,3} \mathbb{P}_\pi(X_1^w \mid COD_j) * \mathbb{P}(COD_j) \; + \; \sum_{j=1,2,3} \mathbb{O}(X_1^w \mid COD_j) * \mathbb{O}(COD_j)} \\
&\quad \overline{+ \mathbb{P}_{\pi_{non}}(X_1^w \mid NON) * \mathbb{P}(NON)}.
\end{aligned}$$

donne la probabilité *a posteriori* de l'événement qu'un fragment donné X_1^w appartienne à une région non-codante.

La somme totale de $\mathbb{P}(COD_f \mid X_1^w)$ et $\mathbb{O}(COD_f \mid X_1^w) f = 1, 2, 3$ est désignée par $\mathbb{P}(COD \mid X_1^w)$. On suppose que $\mathbb{P}(COD \mid X_1^w) + \mathbb{P}(NON \mid X_1^w) = 1$, le cas d'un fragment partiellement codant et

non-codant n'est donc pas pris en compte. Enfin, on se demande si la méthode donne les mêmes résultats de prédiction de CDS qu'on analyse le brin direct ou le brin inverse. La vraisemblance d'un fragment dans le modèle *vers la droite* est la même que celle dans le modèle *vers la gauche* à partir du moment où les paramètres ont été estimés par maximum de vraisemblance sur le même jeu d'apprentissage.

C.2 *Prokov*

Pour plus d'information sur les options de ces programmes en C se référer à l'option -h qui affichera l'aide et pour le principe général se référer à la figure 3.2 p. 97. Nous utilisons les mêmes notations que pour le reste de la thèse qui utilise le concept de phase de la manière suivante : la phase non codante (COD_0), les trois phases de lecture du brin direct (COD_1, COD_2, COD_3) et les trois phases de lecture du brin inverse (COD_{-1}, COD_{-2}, COD_{-3})

C.2.1 *prokov_orf*

Ce programme de prédiction de gènes par signal prend en entrée une séquence nucléique au format Fasta et crée un fichier de CDS (séquences nucléiques avec stop) au format Fasta. Voici un exemple de commande :
« prokov_orf *-l 600 -o BACSU.fna > BACSU_CDS.fna* ».

L'option -l de ce programme est la longueur minimale des CDS que l'on veut sélectionner. Par exemple dans le cadre d'une annotation fine on peut demander -l 60. La longueur minimale de la CDS est 60 pb sans compter la longueur du codon de terminaison. Si on compte le codon de terminaison, alors avec -l 60 on obtiendra des CDS dont la longueur totale (codon de terminaison inclus) sera en fait au minimum de 63 pb. L'option -o permet d'avoir en sortie les données de séquence et non pas juste la liste de CDS et de leurs positions.

Quelque soit le sens direct ou inverse de la CDS, ces positions de début et de fin sont par convention celles projetées sur le brin direct. Quelque soit son sens le début d'une CDS est toujours inférieure à sa fin. La phase d'un fragment d'ADN, $f \in \{-3, -2, -1, 0, 1, 2, 3\}$, ne doit pas être confondue avec la position *pos* d'un oligonucléotide par rapport au codon (voir *prokov_learn*). Plusieurs conventions sont possibles pour le calcul de la phase d'une CDS.

- La convention mathématique est la suivante : $f = ((debut - 1)\%3) + 1$.
- La convention biologique est la suivante :
  - CDS sens direct : $f = ((debut - 1)\%3) + 1$,
  - CDS sens inverse : $f = (((longueur - fin)\%3) + 1) * (-1)$.

La convention mathématique est plus simple mais choque le biologiste qui s'attend à ce que *le premier nucléotide de la séquence inverse complémentaire d'un fragment en phase* -1 *corresponde à la position* 1 *du codon. prokov_orf* utilise la convention mathématique. Nous avons vu que pour les génomes A+T riches, les longues CDS sont avérées. *Prokov* recherche des CDS pour les trois phases positives de la séquence puis pour les trois phases négatives de la séquence complémentaire inverse.

Pour chaque phase, il cherche le premier codon d'initiation rencontré puis le premier codon de terminaison rencontré et crée la première CDS si sa longueur est supérieure au seuil. Puis il cherche le prochain codon d'initiation jusqu'à atteindre la fin de la séquence. Quand on utilise *prokov_orf* pour construire un jeu de CDS qui servira d'apprentissage à *prokov_learn*, il suffit de choisir une longueur minimum des CDS suffisamment grande par exemple 600 pb pour sélectionner les CDS avérées [Hayes & Borodovsky, 1998b]. Dans le cas du génome de *B. subtilis*, on obtient 2689 CDS de longueur minimale (codon de terminaison inclus) 603 pb. Sur la sortie graphique de la figure, on voit que sur les 3000 première paires de base de la séquence de *B. subtilis* parmi toutes les CDS de longueur supérieure à 63 pb (rectangles rouges) seules deux CDS ont une longueur supérieure à 603 pb (rectangles quadrillés rouges en phase +2 et +1).

C.2.2 *prokov_learn*

Ce programme d'apprentissage de gènes par contenu prend en entrée un fichier de CDS (séquences nucléiques avec stop) au format Fasta et crée une matrice des effectifs du m-uple. Voici un exemple de commande :
« prokov_learn *-a -I -k 5 BACSU_CDS.fna BACSU_NC.fna -o BACSU_matNC.txt* ».

L'option -a permet d'avoir une sortie ascii et non pas une sortie binaire (information compréssée). L'option -I permet d'éliminer les codons de terminaison que l'on ne veut pas prendre en compte dans le modèle codant. L'option -o permet d'indiquer le nom du fichier de sortie.

Le modèle codant est un modèle de chaîne de Markov non-homogène (dépendant de la phase ou plutôt de la position du nucléotide dans le codon) du type Mk_3. *On définit la position d'un mot par rapport au codon par la position de son dernier nucléotide dans le codon alors que dans le cas de la phase d'un fragment on se réfère au premier nucléotide.* L'option -k permet de choisir la longueur du k-uple pour le modèle codant. En effet pour apprendre un modèle de chaîne de Markov d'ordre $k-1$, il faut compter les occurrences de tous les k-uplets. Ainsi dans un modèle de chaîne de Markov d'ordre 4, on pourra calculer la probabilité d'apparition d'une base en position i connaissant les bases en position $i-1, i-2, i-3, i-4$ grâce à la matrice de transition où les fréquences d'apparition des tous les mots de longueurs 5 auront été estimées sur un jeu d'apprentissage. Dans notre exemple, on voit que le mot $AAAAA$ apparaît 7335 fois en position 1 des codons, 8169 fois en position 2 des codons et 5179 fois en position 3s des codons de la séquence directe de *B. subtilis*. En théorie il faut compter les effectifs de tous le k-uplets pour les trois positions de codon du brin complémentaire inverse. En pratique, les effectifs du brin complémentaire inverse se déduisent directement des effectifs du brin direct. La relation entre la position pos d'un k-uplet et la position pos' de son inverse complémentaire s'écrit : $pos' = -((3-(k+pos)\%3)\%3)$. Cette relation permet de calculer directement les effectifs des k-uplets sur le brin complémentaire à partir des effectifs sur le brin direct. On en déduit que le mot $TTTTT$ apparaît 7335 fois en position 1 des codons, 5179 fois en position 2 des codons et 8169 fois en position 3 des codons pour la séquence complémentaire inverse de *B. subtilis*. Il est recommandé d'estimer la taille de son jeu d'apprentissage. En effet, pour que les statistiques soient significatives, il faut que chaque k-uplet soit présent au moins cent

fois dans les séquences du jeu d'apprentissage. Pour un modèle d'ordre 4 et un alphabet à 4 lettres, le nombre de 5-uplets est : $N_{5-uplet} = 4^5 = 1024$. Si l'on veut que les 1024 5-uplets soient au moins présents cent fois dans le jeu d'apprentissage, il faut que la taille totale de ce jeu soit au moins égale à $1024 * 100 = 102400$ pb. La taille de notre jeu d'apprentissage BACSU_CDS.fna est de 3220131 ce qui est environ 30 fois supérieur au minimum requis. Nous obtenons ce nombre grâce à la commande suivante : « *grep* $'[A-Z]'$ *BACSU_CDS.fna* | *tr -d* $'\n'$ | *wc -c* ».

Le modèle non-codant est un modèle de chaîne de Markov homogène du type Mk. En général on a moins de séquences non-codantes que de séquences de codantes (le nombre de séquences et la longueur des séquences sont moins importants). Il est donc recommandé de calculer la taille du jeu d'apprentissage des séquences non-codantes et si nécessaire d'utiliser un ordre moins élevé pour la matrice du non-codant. L'option -K permet de choisir la longueur du k-uple pour le modèle non-codant dans le cas où un jeu de séquences non-codantes est spécifié. Dans l'exemple de la figure, il n'y a pas de jeu de séquences non-codantes. Quand aucun jeu de séquences non-codantes n'est spécifié *prokov_learn* simule le modèle non-codant en mélangeant *shuffle* les effectifs des modèles codants. L'effectif d'un k-uplet dans le modèle non-codant est la moyenne des effectifs de ce même k-uplet sur les 6 phases du modèle codant :

$$N_{k-uplet,0} = \frac{1}{6}(N_{k-uplet,1} + N_{k-uplet,2} + N_{k-uplet,3} + N_{k-uplet,-1} + N_{k-uplet,-2} + N_{k-uplet,-3}).$$

Par exemple calculons les effectifs pour le 5-uplet $AAAAA$ dans le modèle non codant :

$$\begin{aligned} N_{AAAAA,0} &= \frac{1}{6}(N_{AAAAA,1} + N_{AAAAA,2} + N_{AAAAA,3} + N_{AAAAA,-1} + N_{AAAAA,-2} + N_{AAAAA,-3}) \\ &= \frac{1}{6}(N_{AAAAA,1} + N_{AAAAA,2} + N_{AAAAA,3} + N_{TTTTT,-1} + N_{TTTTT,-2} + N_{TTTTT,-3}) \\ &= \frac{1}{6}(7335 + 8169 + 5179 + 3620 + 1997 + 3105) = 4900. \end{aligned}$$

Si maintenant on utilise un jeu de séquences non codantes natives de *B. subtilis* alors $N_{AAAAA,0}$ est égale à 1686 mais le taille du jeu de séquences non-codantes est de 504840 pb et non pas de 3220131. Si on calcule le pourcentage de $N_{AAAAA,0}$ par rapport à la taille du jeu, on obtient 0.33% dans le cas du non-codant natif et 0.15% dans le cas du non-codant *shuffle*. Autrement dit il y aurait deux fois plus de $AAAAA$ dans les séquences non-codantes que dans les séquences codantes. Ainsi nous conseillons d'utiliser un jeu d'apprentissage natif pour le modèle non-codant si on en possède un de taille suffisante. A. Viari m'a signalé de ne pas utiliser l'option -K car un petit bogue n'a pas été corrigé. Quand on donne à *prokov_learn* un jeu pour le codant, un jeu pour le non-codant et l'option -k sans préciser l'option -K alors *prokov_learn* utilise le même k pour le non-codant et le codant. Dans ces conditions il n'y a pas de bogue. Ainsi en pratique si le jeu de non-codant natif est trop petit par rapport au jeu de codant, soit on fait du *shuffle* soit on diminue l'ordre du codant.

Le fichier de sortie BACSU_mat.txt contient les matrices du modèle codant (trois colonnes) puis la matrice du modèle non-codant (une colonne). La première ligne de ce fichier donne la version

de *Prokov* (1.2), le k du modèle codant (5), le K du modèle non-codant (5), le nombre de k-uplets pour le modèle codant (1024) et le nombre de K-uplets pour le modèle non-codant (1024).

Si l'on veut utiliser *prokov_learn* et poursuivre par *prokov_score*, cela n'apporte rien de choisir des ordres élevés. Si l'on veut utiliser *prokov_learn* et poursuivre par *prokov_curve*, choisir un ordre élevé améliorera les courbe de prédiction de codage (diminuer le bruit de fond des pics parasites). Lorsqu'on utilise *Prokov* sur des génomes complets de bactéries, on peut se permettre d'utiliser, un ordre du modèle codant élevé et une longueur minimum de CDS importante. Lorsqu'on travaille sur des fragments plus petits, il ne faut pas hésiter à réduire l'ordre du modèle (2-3) voir la longueur minimum des CDS (300-500), les résultats seront statistiquement significatifs et resteront cohérents. Pour vérifier que les effectifs de la matrice sont en nombre suffisant, il suffit d'utiliser l'option -a, et d'aller lire la matrice au format texte pour vérifier qu'en moyenne les effectifs sont bien supérieur à 100.

C.2.3 *prokov_curve*

Ce programme de reconnaissance de gènes par contenu prend en entrée une séquence nucléique au format Fasta et une matrice d'effectifs de k-uple et calcule la probabilité de codage le long des 6 phases de la séquence au moyen d'une fenêtre glissante. Voici un exemple de commande :
prokov_curve *-w 96 -s 12 -m BACSU_mat.bin BACSU.fna > BACSU_curve.txt*

L'option -m permet d'indiquer le nom du fichier des matrices. A partir des tableaux d'occurrences du fichier de modèles de chaîne de markov pour le codant Mk_3 et pour le non-codant Mk, *prokov_curve* commence par calculer respectivement les probabilités de transitions $\pi_{pos}(a_1a_2\dots a_k, b)$ et $\pi(a_1a_2\dots a_k, b)$ et les probabilités initiales $\mu_{pos}(a_1a_2\dots a_k)$ et $\mu(a_1a_2\dots a_k)$. Puis pour chaque phase f et à chaque pas -s sur la séquence, *prokov_curve* calcule la probabilité de codage du fragment X_1^w de taille -w, $\mathbb{P}(COD_f \mid X_1^w)$, grâce à la formule dite d'inversion de Bayes et des matrices de probabilités de transition. Cette probabilité de codage est associée à la position du nucléotide correspondant au milieu du fragment contenu dans la fenêtre glissante. Pour le calcul de la phase c'est la convention biologique qui a été choisie. Il est prévu de changer la convention de phase de *prokov_orf* : choisir comme pour *prokov_curve* la convention biologique. C'est d'ailleurs celle-ci qui a été choisie dans Imagene et *Genostar*. L'option -p permet de changer la probabilité de codage a priori qu'un fragment d'ADN choisi au hasard dans la séquence porte un gène, $\mathbb{P}(COD)$. Comme nous l'avons déjà vu les génomes bactériens sont compacts et on considère généralement que 80% du génome est codant ($\mathbb{P}(COD) = 0.8$ et $\mathbb{P}(NON) = 0.2$). On fait l'hypothèse que les gènes sont uniformément répartis sur les 6 phases. La probabilité a priori d'être codant en phase f, nécessaire à la formule de Bayes, est donc égale à $\mathbb{P}(COD_f) = \mathbb{P}(COD)/6$.

Les résultats de *prokov_curve* sont sensibles à la matrice (ordre et organisme). Dans notre exemple BACSU_curve.txt, pour une fenêtre de taille 96 la première probabilité de codage estimée correspond donc à la position 96/2=49, avec un pas de 12, la seconde correspond donc à la position $49 + 12 = 61$. Pour ces deux positions nous avons 7 valeurs de probabilités, une pour chacun des 6 modèles de phase codante (+1, +2, +3, -1, -2, et -3) et une pour le modèle non-codant (phase

arbitraire 0). La sortie graphique qui représente les courbes de probabilités sur les 6 phases nous montre que pour ces deux positions nous sommes dans une région non-codante : $\mathbb{P}(COD_0 \mid X_{44}^{55}) = 0.795$. En revanche pour la position 1261 correspondant à la CDS en phase +2, on observe une forte probabilité de codage en phase +2 : $\mathbb{P}(COD_2 \mid X_{1256}^{1267}) = 1$.

C.2.4 *prokov_score*

Ce programme de reconnaissance de gènes par contenu prend en entrée un fichier de CDS (séquences nucléiques avec codon de terminaison) au format Fasta et une matrice d'effectifs de m-uple et calcule le score ou probabilité baysienne en appliquant directement la formule de Bayes sur chacune des CDS. Glimmer applique un calcul de score sur chacune des CDS complète mais ce n'est pas la formule de Bayes. Voici un exemple de commande :
« prokov_score *-I -m BACSU_mat.bin BACSU_CDS.fna > BACSU_CDS.score* ».

L'option -m permet d'indiquer le nom du fichier des matrices. L'option -I permet d'éliminer les codons de terminaisons des CDS qui ne doivent pas être pris en compte dans le calcul du score. En utilisant l'option -o qui permet d'obtenir en sortie les données de séquences en plus des scores, j'ai observé que les séquences Fasta étaient raccourcies non pas des 3 dernières bases mais des 4 dernières bases, ce peut être un bogue.

A partir des tableaux d'occurrences du fichier de modèles de chaîne de Markov pour le codant Mm_3 et pour le non-codant Mm, *prokov_score* commence par calculer les probabilités de transitions et les probabilités initiales. Puis pour le calcul des scores, on obtient pour chaque CDS sept scores, un pour chacune des phases codantes et un pour la phase non-codante. Ce score est la probabilité de codage en phase f de la CDS $\mathbb{P}(COD_f \mid CDS)$. Pour le calcul de la phase c'est la convention biologique qui a été choisie. Si la CDS est avérée alors elle a un score élevé en phase +1. $\mathbb{P}(COD_1 \mid CDS)$ correspond à la deuxième colonne du fichier de scores, la première correspondant au nom de CDS. Peu importe que la CDS soit dans le sens direct ou inverse, si elle est avérée alors $\mathbb{P}(COD_1 \mid CDS)$ est élevée. En effet les CDS sont toutes décrites en phase +1 (le premier nucléotide de la séquence correspond à la première position du codon d'initiation). Si la CDS est non-codante, elle aura un score élevé en phase 0 (dernière colonne du fichier de sortie). Si la CDS possède un décalage du cadre de lecture, elle aura un score élevé dans la phase +2 ou +3. Les résultats de *prokov_score* sont peu sensibles à la matrice (ordre et organisme). Les valeurs de score des CDS avérées ont tendance à être très élevés, elles sont souvent égales à 1. Un score de 99.9, signifie que l'on a une chance sur 1000 de se tromper en prédisant cette séquence comme codante en phase +1. Ce score est donc puissant car il arrive à prédire sans se tromper, mais il est peu sensible car toutes les CDS avérées ont un score élevé proche de 1 (un peu comme si c'était noir ou blanc sans les dégradés de gris).

C.2.5 *prokov_cds*

Ce programme de prédiction de gènes par signal et par contenu prend en entrée une séquence nucléique au format Fasta et une matrice d'effectifs de m-uple, d'abord il crée un fichier de CDS puis il calcule les probabilités de transitions et les probabilités initiales des matrices et enfin il calcule le score dans la phase +1 associé à chaque CDS. Voici un exemple de commande :
« prokov_cds *-l 60 -m BACSU_mat.bin BACSU.fna > BACSU_CDS_score.lst* ».

L'option -l correspond à la longueur minimale des CDS. L'option -m permet d'indiquer le nom du fichier des matrices. Il n'y pas d'option -I, avec l'option -o les CDS sont bien affichées avec les codons de terminaison. Cependant à l'intérieur du programme, le calcul du score ignore le codon de terminaison, la gestion des codons de terminaison se fait en interne. Il n'y a pas de calcul de phase, puisque seul le score de la phase +1 de la CDS est calculé. Les options -w et -W sont sensées regarder si en prenant un codon d'initiation en 3', on obtient un meilleur score et dans ce cas changer le codon d'initiation. Elles ne doivent pas être utilisées car elles ne fonctionnent pas bien.

Finalement *prokov_cds* c'est l'équivalent de *prokov_orf* suivie de *prokov_score*. Le problème du choix du codon d'initiation reste non résolu.

Annexe D

Méthodes d'analyse factorielle

D.1 Principe de l'analyse en composantes principales (ACP)

Les principales d'une ACP usuelle sont les suivantes :

1. Le tableau de données X, à n individus et à p variables, est d'abord transformé en une matrice Z centrée réduite suivant une métrique euclidienne *classique* afin de supprimer l'hétérogénéité des variables (qui sont en général mesurées avec des unités différentes). Une métrique M est une matrice qui permet de définir un produit scalaire et donc des distances[1] entre individus ou entre variables. En ACP *non normée* on utilise la métrique diagonale où les p éléments de la diagonale on pour valeur $1/s^2$. La distance entre deux individus se calcule donc : $d^2(i,i') = \sum_j 1/s_j^2(x_{ij} - x_{i'j})^2$. On montre qu'utiliser cette métrique revient à multiplier chaque variable par $\sqrt{1/s_1^2}$ (*e.g.* à diviser chaque variable par son écart type) et à utiliser la métrique identité $M = I$. La réduction des données de X permet donc d'utiliser la métrique identité ($ZM = Z$) et de réaliser une ACP *normée*. En réalité dans l'ACP, aucune matrice de distances entre les points pris deux à deux n'est calculée. Cependant, la métrique euclidienne est à la base du calcul de la matrice Z du nuage centré de points $N(I)$ dans un espace à p dimensions.
2. Une matrice R de variance – covariance des p variables est ensuite déduite de cette matrice $Z.R$ est une matrice carrée de dimension p. On montre que cette matrice R est la matrice d'inertie du nuage de points $N(I)$. L'inertie totale (la trace[2]) de R est égale à np.
3. L'étape centrale du calcul de l'ACP est la diagonalisation de la matrice d'inertie qui consiste à chercher les axes de symétrie de l'ellipsoïde d'inertie associé au nuage de points $N(I)$. La matrice R diagonalisée est réduite à une diagonale dont les éléments s'appelle les valeurs propres λ. La diagonalisation n'a pas changé la trace de R mais à supprimé tous les termes de covariance. Puis les vecteurs propres u de la matrice R sont déduits par $Ru = \lambda u$. Chaque *axe principal* k est donc défini par un vecteur propre u_k de la matrice R et caractérisé par

[1] Un indice de distance d(i,j) vérifie les propriétés suivantes : $d(i,j) \geq 0$; $d(i,j) = d(j,i)$; $d(i,i) = 0$. Une distance euclidienne respecte l'inégalité triangulaire : $d(i,j) \leq d(i,k) + d(k,j)$.

[2] La trace est la somme des valeurs diagonales d'une matrice carrée

une valeur propre (variance) λ_k de cette matrice. Les coordonnées des points des individus résultent des projections orthogonales de Z sur les *axes principaux*. Ainsi sont obtenues n *composantes principales* c contenant les coordonnées des individus sur p *axes principaux*. Finalement, on s'aperçoit que les *facteurs principaux* f (de la première définition) sont définis tels que $f = Mu$, ce qui signifie dans la cas de la matrice identité que les facteurs principaux sont les vecteurs propres de R ($f = u$). On retrouve donc l'idée que les composantes principales sont définies par les combinaisons linéaires des variables initiales de variance maximale ($c_{ik} = Zf_k = Zu_k$).

De la même manière, il est possible de réaliser une ACP sur le nuage de points des variables. Cependant, l'interprétation des projections du nuage des individus N(I) ne fait pas référence à celle du nuage des variables N(J). En effet, les coordonnées n'ont pas été calculées selon la même base orthonormée : on projette le nuage des variables sur les n *axes factoriels* (et non sur les p *axes principaux*). En pratique on construit les projections des nuages $N(I)$ et $N(J)$ en procédant à *une seule* ACP [Chiapello, 1999].

D.2 L'analyse factorielle des correspondances (AFC)

D.2.1 Méthode de l'AFC

D'un point de vue mathématique, les principales étapes d'une AFC sont les suivantes (Fig. 3.4 p. 126) :

1. A partir du tableau de données X, à n individus et à p variables, on construit un tableau de fréquences (ou contributions) relatives F en divisant chaque case du tableau de données par son total général T. On peut associer à ce tableau de fréquences *conjointes* de terme général $f_{ij} = x_{ij}/T$ le nuage des n profils des lignes de coordonnée courante f_{ij}/f_i et le nuage des p profils des colonnes de coordonnée courante f_{ij}/f_j. En pratique, il n'est pas nécessaire de calculer ni le tableau des profils des lignes ni celui des profils des colonnes ; cependant la métrique du χ^2 utilisée à l'étape suivante repose sur cette notion de profils qui pondère la fréquence *conjointe* par la fréquence *marginale* ($f_i = x_i/T$ ou $f_j = x_j/T$) donnant ainsi plus de poids aux écarts correspondants à des effectifs faibles (*e.g.* un gène de vingt codons aura le même poids qu'un gène de trois cent [Baccini & Besse, 2002]).
2. La matrice intermédiaire Y est calculée à partir du tableau F, selon la métrique du χ^2. La matrice Y a pour terme courant $y_{ij} = (f_{ij} - f_i f_j)/\sqrt{f_i f_j}$ qui correspond à la racine carrée du χ^2. Ici on calcule un indice de distance entre une ligne (*e.g.* un gène) et le centre de gravité du nuage (*e.g.* usage des codons moyen dans l'échantillon). La métrique du χ^2 (tout comme la métrique euclidienne classique de l'ACP usuelle) permet de définir une distance entre deux points. La distance entre deux profils des lignes est $d^2(i,i') = \sum_{j=1}^{p}(f_{ij}/f_i - f_{i'j}/f_{i'})^2/f_j$. De même, la distance entre deux profils des colonnes, est $d^2(j,j') = \sum_{i=1}^{n}(f_{ij}/f_j - f_{ij'}/f_{j'})^2/f_i$.

3. La matrice R d'inertie entre colonnes est ensuite déduite de cette matrice Y comme pour l'ACP usuelle.
4. De même, la diagonalisation de la matrice R permet de définir les axes de projection (vecteurs propres) et de les classer suivant leur pourcentage d'inertie décroissant (valeurs propres). Les coordonnées des profils des lignes sur les axes de projections sont calculées à partir des valeurs propres et des coefficients des vecteurs propres. En AFC, axe principal est synonyme d'axe factoriel. Enfin, d'autres mesures sont calculées pour aider à l'interprétation des résultats.

En pratique comme dans le cas de l'ACP usuelle, on construit les projections des nuages $N(I)$ et $N(J)$ en procédant à *une seule* ACP. On calcule la matrice d'inertie de plus petite dimension par exemple la matrice d'inertie entre colonnes si $p < n$ correspondant à l'ACP du nuage des profils des lignes. Finalement, l'AFC peut être considérée comme un cas particulier de l'ACP. En effet, dans la théorie de l'ACP, on peut choisir la masse de chaque point (*i.e.* la pondération) et la distance entre les points (*i.e.* la métrique). Dans une AFC, les points sont des profils, c'est-à-dire que leurs composantes ne correspondent pas à des valeurs numériques observées mais à des fréquences d'apparition dans la population totale. La masse d'un point est donc imposée, tout comme la distance entre deux points : la métrique du χ^2 respecte la notion de profil, et donc de fréquences relatives. L'originalité de l'AFC est de permettre la représentation simultanée des deux nuages des profils (des lignes et des colonnes) sur le même graphique. Cette opération est possible car il existe des relations simples entre les coordonnées factorielles des deux nuages qui découlent des conditions imposées à l'AFC [Chiapello, 1999].

D.2.2 Le programme AFCcodons

Nous allons décrire l'utilisation du programme *AFCcodons* dont les sources C nous ont aimablement été fournis par Hélène Chiapello (Fig. 3.5 p. 128).

Fichier et paramètres d'entrée

Le programme prend en entrée une liste de CDS au forma Fasta. Le programme attend un seul champ derrière le chevron supérieur marquant le début d'une CDS. Si l'entête Fasta contient plusieurs champs le programme ne produit pas d'erreur mais les résultats statistiques de l'AFC sont faux. Si le champ d'identification de la CDS fait plus de six caractères, les caractères supplémentaires sont tronqués dans le fichier des résultats statistiques de l'AFC. Le fichier BACSU_CDS.fna contient 2689 CDS de longueur supérieure ou égale à 603 pb codon de terminaison inclus de BACSU. On visualise graphiquement la phase -3 de la région qui va de 3634000 à 3660000.

Les CDS représentées par de rectangles rouges ont une longueur supérieure ou égale à 63 pb codon de terminaison inclus. Les CDS représentées par des rectangles rouges quadrillées sont annotées dans le fichier de WWDDL. Au cours de cet exemple nous allons suivre plus particulièrement les CDS 3634127C, 3650127C et 3658149C de longueur supérieure ou égale à 603 pb codon de terminaison inclus et présentes dans le fichier BACSU_CDS.fna. On remarque que la prédiction de codage

(courbe bleue) est meilleure pour les CDS 3634127C et 3650127C que pour la CDS 3658149C, qu'en est il de leur usage des codons synonymes ? Le programme *AFCcodons* a besoin de deux paramètres la longueur minimum des CDS et le code génétique. La longueur des CDS est réglée à 204 pb codon de terminaison inclus (longueur minimale pour que les résultats statistiques de l'AFC soient significatifs). On choisit le code génétique bactérien (choix numéro 10). Il correspond au fichier transl_table_11.txt.

Tableau des fréquences relatives des codons synonymes

Le programme commence par vérifier chacune des séquences. Les problèmes mis en évidence dans les CDS (longueur insuffisante, non multiple de trois, le dernier codon n'est pas un codon de terminaison, le premier codon n'est pas un codon d'initiation, la séquence possède un codon de terminaison à telle position) sont répertoriés dans le fichier.seqpb. Le programme compte les effectifs des codons dans les gènes. Le tableau des effectifs de codons des gènes est stocké dans le fichier.eff. L'extrait du fichier.eff choisi décrit les effectifs des codons TTA, ATA et ATC codant le même acide aminé isoleucine pour les trois gènes 3634127C, 3650127C et 3658149C. A partir des effectifs, le programme calcule les RSCU. Le tableau des RSCU des gènes est stocké dans le fichier.data. Les lignes correspondent aux CDS et les colonnes aux codons. Analysons l'extrait du fichier.data. Pour 3650127C, le codon préféré pour l'isoleucine est ATT (56,25%) et le codon détesté est ATA (0%). Pour 3634017C, le codon préféré pour l'isoleucine est ATC (54,55%) et le codon détesté est ATA (6,25%). C'est la CDS qui présente un usage des codons synonymes le plus biaisé (54,55 – 0 = 54,55). Pour 3658149C, le codon préféré pour l'isoleucine est ATT (57,14%) comme pour 3650127C. Cependant le codon ATA n'est pas détesté (22,86%). C'est donc la CDS la moins biaisée (57,14 – 20 = 37,14).

Matrice d'inertie du nuage de codons

Le programme transforme le tableau RSCU en un tableau de fréquences. Ce tableau de fréquences représente deux profils, le profil des lignes et le profil des colonnes, qui peuvent être vus comme deux nuages de points le nuage des gènes et le nuage des codons. Chaque case du tableau est divisé par le total général. Le total général est la somme de toutes les valeurs RSCU du tableau. Si on somme les valeurs pour chaque ligne et dans le cas particulier de l'usage des codons synonymes, la somme des RSCU pour les codons d'un même acide aminé vaut un. La somme des valeurs d'une ligne vaut donc le nombre d'acide aminé, au plus 18. Le total général est donc le produit du nombre de gènes par le nombre d'acides aminés. Ici il vaut 45713 soit le produit de 4289 CDS par 17 acides aminés.

Pour le nuage de gènes, chaque gène est représenté par un point dans l'espace des codons à 57 dimensions. Par rapport au 59 codons généralement étudiés (64 moins le codon de la méthionine, le codon du tryptophane et les trois codons de terminaison), nous avons éliminé les deux codons cystéines car ce sont des codons rares, ce cas de figure n'est pas souhaité. La coordonnée du gène

i sur l'axe j représente la fréquence relative du codon j dans le gène i. Pour le nuage de codons, chaque codon est représenté par un point dans l'espace des gènes à 2689 dimensions. La coordonnée du codon j sur l'axe i représente la fréquence relative du codon j dans le gène i. Nous aimerions connaître l'organisation de ces points dans l'espace, et les caractéristiques remarquables qui peuvent en être déduites. La quantité de données ne permet pas d'interpréter les profils par la lecture seule du contenu du tableau de fréquences. L'idée est de mettre en évidence les tendances des profils par une visualisation graphique des déformations du nuage de points (s'il n'y a aucun biais le nuage est sphérique). Le cerveau humain n'est pas apte à se représenter efficacement des données définies par plus de trois dimensions. En effet pour visualiser les corrélations entre p variables prises deux à deux et mesurées sur n individus, il faudrait réaliser p(p-1)/2 graphiques dont l'analyse serait fastidieuse.

L'astuce apportée par les techniques d'analyse de données multivariées des statistiques descriptives consiste à trouver les axes de symétrie du nuage de points qui caractérisent le mieux les tendances générales du nuage. Ces axes doivent être orthogonaux de manière à mettre en évidence des tendances qui sont indépendantes les unes des autres. L'information portée par ces nouveaux axes ne sera donc pas corrélée. On peut ainsi dégager les principaux facteurs qui déforment le nuage de manière indépendante. L'origine de ce nouveau repère est le centre de gravité du nuage de points. Pour trouver ce nouveau repère on va se baser non pas sur les coordonnées des points dans l'ancien repère, mais sur les distances entre les points pris deux à deux. Cette matrice de distance des points pris deux à deux est appelée matrice d'inertie R. C'est une matrice carrée et symétrique dont la dimension est le nombre de points.

En pratique on verra que connaissant les coordonnées des points du nuage de codons dans le nouveau repère on peut facilement calculer celles des points du nuage de gènes dans ce même repère. C'est l'avantage majeur de l'AFC pouvoir superposer dans un même espace le nuage des individus (les gènes) et celui des variables (les codons). A ce stade il nous suffit donc de calculer une seule des deux matrices d'inertie et de préférence celle de dimension la plus petite. La matrice d'inertie des codons est de taille 59 sur 59. Dans notre exemple la matrice d'inertie des gènes est de taille 2689 sur 2689. On choisit donc de calculer la matrice d'inertie des codons. La distance choisie est celle du χ^2. Le programme calcule d'abord la matrice intermédiaire Y, de distances de χ^2 entre chaque point et le barycentre du nuage de points.

Le programme déduit la matrice d'inertie R à partir de la matrice intermédiaire et de sa transposée (R = Y * YT). Ainsi nous obtenons la matrice d'inertie R, de distances de χ^2 entre chaque codon pris deux à deux.

Diagonalisation de la matrice d'inertie

C'est en diagonalisant la matrice R ou matrice de variance/covariance que l'on va pouvoir accéder aux vecteurs propres U (axe de symétrie) et aux valeurs propres λ (mesure de l'inertie portée par les axes). La diagonalisation de la matrice d'inertie est l'étape clé de l'AFC ou de l'ACP. La diagonalisation de la matrice d'inertie consiste à rechercher les axes de symétries de

l'ellipsoïde d'inertie associé au nuage de codons et à mesurer l'inertie portée par chacun d'entre eux. La matrice R diagonalisée est réduite à une diagonale dont les éléments s'appellent les valeurs propres. La diagonalisation n'a pas changé la variance de R (l'inertie totale ou la trace de R) mais a annulé tous les termes de covariance. A chaque valeur propre λ de la matrice diagonalisée est associé un vecteur propre non nul tel que $R_U = \lambda_U$. Ainsi, les coordonnées des vecteurs propres sur les axes factoriels s'obtiennent par la résolution d'équations matricielles. En pratique l'axe de symétrie de plus grande inertie associé au nuage de codons n'est autre que la tendance dominante des données, c'est-à-dire la direction dans laquelle les points sont le plus dispersés (s'écartent le plus de la moyenne). Les axes sont donc classés dans l'ordre décroissant de valeurs propres. L'inertie totale et les pourcentages d'inertie (valeur propre ou variance) de chacun des axes du nuage de codons sont stockés dans le fichier.out. Dans notre exemple l'axe 1 et l'axe 2 porte 12 et 11% de l'information respectivement. 56 axes ont été déterminés car nous n'avons que 56 variables indépendantes.

Projection des codons sur les axes des vecteurs propres

Les coordonnées des points du nuage dans ce nouveau système d'axes sont calculées par projection orthogonale des points sur les axes de projection (axe de symétrie de l'ellipsoïde).

Les coordonnées des vecteurs propres définissent les axes de projection des nuages N(I) et N(J). La projection des points colonnes sur les axes de projection s'obtient grâce au produit scalaire des valeurs propres par les composantes des vecteurs propres. Plus précisément la coordonnée x_{jk} de la projection du point colonne j suivant l'axe k dépend de u_{jk}, λ_k et f_j (total de colonne j dans la matrice F) $x_{jk} = u_{jk} \times \sqrt{\lambda_k / f_j}$. Le programme ne calcule les coordonnées de projection des points que pour les axes portant au moins 2% de l'information. En dessous de ce seuil l'information portée par les axes n'est pas significative. Dans notre exemple, le programme calcule les coordonnées des projection des points colonnes pour les 17 premiers axes sont les seuls à porter plus de 2% d'inertie. Le programme calcule aussi pour les coordonnées de projection des codons sur les 5 premiers axes deux indices d'aide à l'interprétation : les cosinus carrés (rho) et les contributions absolues (ct ; Tab. D.1 p. 405) de chaque variable aux axes de projection. Calculer ces indices sur les 5 premiers axes est amplement suffisant, car on n'examine en général pas plus de trois ou quatre axes de projection. Les intitulés des points colonnes (codons) sont stockés dans le fichier_col.row. Les coordonnés des projection des codons sont stockées dans le fichier_col.dat. Pour alléger le texte, nous écrirons par la suite les coordonnées des points au lieu du terme exact les coordonnées des projections des points. Les cosinus carrés et les contributions de chaque codon aux axes sont stockés à la suite des résultats d'inertie dans le fichier.out. Le type de points des codons pour la visualisation graphique avec xgobi est stocké dans le fichier_col.glyph.

Projection des gènes sur les axes des vecteurs propres

La projection des points individus sur les axes de projection s'obtient grâce à la relation de dualité qui se définie ainsi :

	BACSU				ECOLI				MYCTU				PHOLU			
item	axe1	ct*100	axe2	ct*100	axe1	ct*100	axe2	ct*100	axe1	ct*100	axe2	ct*100	axe1	ct*100	axe2	ct*100
AAA_K	-4	28	4	33	-1	1	1	6	-26	**396**	-74	**4376**	0	1	-3	27
AAC_N	5	19	26	**637**	-27	**437**	5	43	14	**350**	-4	44	30	**407**	-19	**317**
AAG_K	9	49	-11	76	4	6	-4	12	13	**281**	26	**1429**	2	1	11	78
AAT_N	-4	21	-19	**478**	31	**510**	-5	32	-47	**1082**	12	**105**	-15	**213**	8	**134**
ACA_T	-5	19	5	24	53	**429**	19	**124**	-29	**153**	-20	**101**	-27	**275**	-9	58
ACC_T	19	**109**	-13	59	-21	**217**	-5	24	12	**209**	-1	7	29	**315**	-7	38
ACG_T	29	**385**	-3	9	6	13	-27	**472**	-11	80	10	92	14	47	29	**389**
ACT_T	-53	**847**	5	12	0	0	40	**654**	-34	**172**	2	1	-10	42	-2	6
AGA_R	-16	**133**	-17	**179**	122	**754**	83	**749**	-46	95	-14	12	-75	**1243**	-30	**365**
AGC_S	17	**114**	4	9	-13	57	-12	**108**	4	9	-7	41	29	**208**	-8	29
AGG_R	0	0	-50	**512**	107	**300**	58	**193**	-22	53	2	1	-45	**190**	-5	5
AGT_S	-35	**239**	-25	**153**	29	**143**	-11	45	-25	93	8	13	-6	12	12	91
ATA_I	-28	**199**	-41	**496**	89	**657**	33	**196**	-64	**502**	-8	12	-37	**364**	-14	**101**
ATC_I	10	76	18	**268**	-24	**265**	7	49	9	**155**	-1	8	32	**393**	-7	35
ATT_I	0	0	-1	4	6	24	-10	**123**	-26	**242**	13	93	-4	12	9	**116**
CAA_Q	-21	**423**	4	19	24	**239**	3	10	-21	**272**	-25	**496**	-13	**138**	-18	**480**
CAC_H	8	38	41	**1108**	-29	**382**	21	**451**	10	**171**	-4	41	35	**498**	-39	**1132**
CAG_Q	24	**485**	-3	15	-13	**135**	-2	6	7	86	9	**189**	18	**199**	23	**600**
CAT_H	1	1	-20	**591**	19	**232**	-19	**489**	-27	**469**	7	51	-13	**165**	18	**555**
CCA_P	-46	**762**	7	25	24	**127**	15	**111**	-24	**149**	-15	79	-14	90	-16	**208**
CCC_P	13	31	-36	**273**	34	**166**	-19	**119**	8	49	-9	82	21	92	-18	**118**
CCG_P	36	**912**	0	0	-31	**523**	-8	81	3	13	9	**141**	30	**309**	33	**680**
CCT_P	-22	**247**	7	30	35	**229**	19	**153**	-38	**202**	-5	7	-20	**170**	-1	3
CGA_R	-10	18	-26	**144**	57	**242**	-3	2	-21	**100**	-10	31	-10	15	-5	7
CGC_R	19	**124**	40	**676**	-18	**144**	-17	**270**	12	**130**	-5	40	36	**353**	-13	94
CGG_R	36	**352**	-21	**150**	31	99	-35	**272**	0	0	8	79	27	**129**	27	**242**
CGT_R	-19	**111**	44	**729**	-21	**181**	12	**141**	-8	19	5	12	16	**132**	14	**180**
CTA_L	-41	**162**	-7	6	46	86	3	1	-18	36	-15	35	-13	21	-28	**170**
CTC_L	25	**121**	1	0	0	0	-8	17	0	0	-11	74	17	29	-21	85
CTG_L	31	**384**	-1	1	-29	**463**	1	2	12	**159**	1	5	43	**614**	2	4
CTT_L	-3	4	16	**144**	31	**120**	10	26	-32	**141**	-3	3	-15	47	-8	25
GAA_E	-4	25	4	28	-1	2	5	55	-8	52	-12	**173**	-3	9	-6	81
GAC_D	8	45	15	**168**	-19	**157**	8	59	9	**131**	-2	15	32	**322**	-20	**230**
GAG_E	8	43	-8	50	3	5	-13	**123**	4	29	6	90	6	19	16	**190**
GAT_D	-5	28	-8	92	11	88	-4	27	-23	**334**	7	47	-9	99	5	68
GCA_A	-17	**154**	-1	2	16	64	19	**197**	-16	59	-11	39	-15	98	-9	62
GCC_A	22	**177**	-4	8	-1	1	-13	**113**	9	85	-3	20	26	**199**	-14	**109**
GCG_A	27	**319**	4	10	-11	50	-19	**292**	-1	2	8	77	18	93	26	**341**
GCT_A	-25	**292**	1	2	5	5	35	**504**	-26	**141**	1	0	-15	92	1	3
GGA_G	-7	31	-4	13	51	**351**	2	3	-16	65	-8	24	-31	**274**	-9	46
GGC_G	19	**226**	14	**135**	-17	**134**	-8	58	9	99	-3	20	28	**262**	-13	**114**
GGG_G	10	30	-25	**228**	20	73	-22	**177**	-7	25	8	43	2	1	22	**216**
GGT_G	-32	**353**	5	14	-8	23	19	**280**	-9	37	6	27	-1	2	3	14
GTA_V	-26	**242**	5	11	7	11	23	**208**	-21	62	-12	25	-18	98	-11	74
GTC_V	26	**292**	2	4	4	5	-11	69	6	39	-6	44	21	**111**	-12	73
GTG_V	20	**197**	-7	31	-10	45	-23	**438**	2	7	7	77	23	**168**	26	**428**
GTT_V	-24	**298**	1	1	6	12	27	**456**	-29	**193**	0	0	-14	**104**	-4	16
TAC_Y	2	4	25	**489**	-25	**298**	12	**138**	17	**464**	-12	**342**	25	**243**	-23	**384**
TAT_Y	0	0	-14	**303**	19	**227**	-9	**118**	-36	**861**	29	**792**	-11	**126**	9	**152**
TCA_S	-6	16	0	0	41	**231**	13	53	-24	92	-8	15	-20	**116**	-10	52
TCC_S	22	**111**	1	0	-25	**105**	15	79	7	27	-3	9	22	75	-12	45
TCG_S	26	**118**	-19	77	-1	0	-31	**328**	4	13	9	98	8	8	22	**101**
TCT_S	-20	**155**	17	**127**	-10	19	41	**583**	-31	95	-7	8	-13	52	0	0
TTA_L	-33	**411**	-7	27	47	**345**	5	11	-38	59	-2	0	-32	**372**	-5	17
TTC_F	-5	14	36	**876**	-26	**305**	19	**364**	12	**257**	-4	41	23	**225**	-19	**284**
TTG_L	-1	0	-11	48	21	66	-12	47	-15	95	14	**111**	0	0	27	**401**
TTT_F	2	5	-16	**395**	17	**197**	-13	**240**	-41	**805**	15	**161**	-10	**107**	8	**129**

TAB. D.1 – Contribution des codons à la formation des deux premiers axes de l'AFC

Contribution (ct*100) des 57 codons synonymes sur les deux premiers axes de projection de l'AFC calculé par *AFCcodon*. Les codons sont triés par ordre alphapétique. Les plus fortes contributions (ct*100>100) sont indiqués en gras.

Abréviations : Axis_nb, nombre d'axes pris en compte par Ether ; Inertia%, pourcentage d'inertie minimum des axes pris en compte, CDS_nb, nombre de CDS.

$$x_{ik} = 1/\sqrt{\lambda_k} \times \sum_{j=1} p(f_{ij}/f_i \times x_{jk})$$
$$\text{où } x_{jk} = 1/\sqrt{\lambda_k} \times \sum_{i=1}^{p} (f_{ij}/f_j \times x_{ik})$$

Le programme calcule aussi les corrélations et les contributions de chaque variable avec les axes de projection. Les intitulés des points lignes (CDS) sont stockés dans le fichier_row.row. Les coordonnés des CDS sont stockées dans le fichier_row.dat. Les cosinus carrés et les contributions de chaque CDS aux axes sont stockés à la suite des résultats d'indices de codons dans le fichier.out. Le type de points des CDS pour la visualisation graphique avec Xgobi est stocké dans le fichier_row.glyph.

Représentation graphique

On peut visualiser les nuages par projection des points sur les plans définis par les couples de vecteurs propres. L'AFC autorise la représentation simultanée sur un même graphique des points lignes et des points colonnes. Ceci facilite l'interprétation des résultats. L'origine de ce graphique représente le centre de gravité des deux nuages de points des codons et des gènes (c'est-à-dire la moyenne pondérée de deux nuages N(I) et N(J)). Ici on a représenté les points selon leurs coordonnées sur les axes 1 et 2. Par exemple, on visualise la CDS 3634017C de coordonnées -37,14 et -56,83. Sur le même graphique on visualise aussi le codon ATA de coordonnées -29,87 et 35,59. L'axe 1 portant 12% de la variance sépare les codons se terminant par T ou A plus fréquents dans les gènes comme 3658149C des codons se terminant par G ou C plus fréquents dans les gènes comme 3650127C. Si seule l'information portée par l'axe 1 est prise en compte, ceci signifie que les codons se terminant par T ou A se comportent de la même manière dans notre échantillon et de même pour les codons se terminant par G ou C. De la même façon les gènes proches de 3658149C présentent un usage des codons similaire ainsi que les gènes proches de 3650127C. L'axe 2 portant 11% de la variance sépare les codons se terminant par T ou C plus fréquents dans les gènes comme 3634017C des codons se terminant par G ou A. Nous avons vu que les codons se terminant entre autre par A sont plus fréquents dans les gènes comme 3658149C alors que les codons se terminant entre autre par G sont plus fréquents dans les gènes comme 3650127C. Ainsi cette superposition graphique, nous aide à mettre en évidence trois tendances qui lient les deux nuages : les gènes comme 3658149C qui préfèrent les codons se terminant par T ou A comme ATA (facteur de dispersion de l'axe 1), les gènes comme 3634017C qui préfèrent les codons se terminant par C ou T comme ATC (facteur de dispersion de l'axe 2). Ces deux tendances minoritaires s'opposent à la tendance majoritaire représentée par les gènes comme 3650127C qui préfèrent les codons se terminant par G. Finalement ces trois CDS reflètent trois usages des codons synonymes distincts.

Pour résumer, les résultats du programme *AFCcodons* sont stockés dans dix fichiers de sortie. Plus généralement, l'AFC appliquée à un tableau de données va rechercher les proximités entre individus et variables basées sur la distance du χ^2. Les points proches dans l'espace se comportent de la même manière. L'AFC va regrouper les variables (les codons) qui ont un profil semblable (ils sont utilisés préférentiellement par le même type de gènes) et regrouper les individus (les CDS) qui ont un profil d'usage des codons synonymes semblable. L'AFC va trier ces groupes selon le critère de dispersion des points. Plus la dispersion est grande, plus le biais est important, et plus le facteur

commun à ce groupe de points joue un rôle prépondérant dans la déformation de l'ellipsoïde. Ainsi l'information est synthétisée (regroupement sur un seul axe de variables ou d'individus ayant des comportements similaires) et ordonnée (classement des axes selon leur inertie ou variance). Ceci nous aide à mettre le doigt sur les principaux biais d'usage des codons synonymes présents dans les jeux de CDS des génomes bactériens.

Annexe E

Classification automatique

E.1 Partition K-means

Le programme utilisé est celui du package cclust du logiciel R (Fig. 3.7 p. 140). Il utilise l'algorithme de Hartigan et Wong 1979 [Hartigan & Wong, 1979]. Le fichier d'entrée est contient n individus, représentés par des points dans un espace euclidien à p dimension. Dans notre cas ce sont les coordonnées de projection sur les 17 axes factoriels d'inertie supérieure à 2 % (information significative voir la règle du pouce dans poly besse) de 2689 gènes de longueur supérieure à 600 pb du chromosome de BACSU. Il prend deux paramètres en entrée le nombre de groupe et le nombre d'itération. Ici on demande 3 groupes et 200 itérations au maximum. Les 3 centres de gravité initiaux sont tirés au hasard. Les points sont affectés au centre de gravité dont ils sont le plus proche. Les centres de gravité des trois groupes ainsi définis sont recalculés. L'algorithme itère jusqu'à stabilisation des groupes. Le programme donne en sortie le numéro du groupe 1,2 ou 3 de chacun des gènes. La représentation graphique du nuage de gènes projeté sur les deux premiers axes d'inertie permet de coloriser les points en fonction de leur groupe. Les points appartenant au groupe 1 sont en noir. Les points appartenant au groupe 2 sont en rouge. Les points appartenant au groupe 3 sont en vert. Si on regarde le pourcentage de contribution en fonction des groupes, on s'aperçoit que les gènes qui ont le plus contribué à la formation de l'axe 1 sont les gènes de groupe 3 dont les codons se terminent préférentiellement par les nucléotides T ou A. En revanche, les gènes qui ont le plus contribué à la formation de l'axe 2 sont les gènes de groupe 2 dont les codons se terminent préférentiellement par les nucléotides C ou T. Ainsi les résultats de l'AFC et de cette classification permettent d'émettre des hypothèses quant à une utilisation différentielle dans l'usage des codons synonymes d'un ensemble de gènes.

E.2 Classification hiérarchique ascendante

Elle consiste à fournir un ensemble de partitions de E en classes de moins en moins fine (ascendante) par regroupements successifs de parties (hiérarchie). Elle se représente par un dendogramme

ou arbre de classification. On associe au système de classes résultant une échelle de niveau : à chaque partition on associe une valeur numérique représentant le niveau auquel on lieu les regroupements (classification hiérarchique indicée). Il existe différentes méthodes selon la stratégie de regroupement. Etant donné un tableau de distances (ou d'indices de distance) d(ij) entre individus, on définit des indices d'agglomération (souvent appelé improprement distances) de deux parties A et B d'un ensemble E. Le poids des parties A et B, somme de poids de leur éléments, est noté P_A et P_B. Leur barycentres sont notés g_A et g_B. Les stratégies d'agrégation définies par les indices d'agglomération sont les suivantes :

1. Soient d la distance initiale calculée sur l'ensemble E $d_{min}(A,B) = inf(d_{ii'},\ i \in A,\ i' \in B)$ (saut minimum ou « single linkage »).
2. $d_{max}(A,B) = sup(d_{ii'},\ i \in A,\ i' \in B)$ (saut maximum ou diamètre de la réunion ou « complete linkage »).
3. $d_{moy}(A,B) = (\sum(d_{ii'},\ i \in A,\ i' \in B))/(cardA \times cardB)$[1] (moyenne des distances ou saut moyen ou « average linkage »).
4. $d(A,B) = d(g_A, g_B)$ (distance des barycentres, critère de Ward pour distances euclidienne).

Remarques :

1. La stratégie du saut minimum permet de reconstituer des classes filiformes mais elles contiennent parfois des éléments très éloignés.
2. Au contraire la méthode du diamètre évite des classes comprenant des éléments trop éloignés. Elle réduit le diamètre des groupes.

[1] Le cardinal d'un ensemble fini (i.e. ayant un nombre fini d'éléments) est son nombre d'éléments.

Annexe F

Objets Génomiques

Field	Type	Null Default	Description of a Genomic Object	Exemple
GO_id	int(11) unsigned	NOT NULL auto_increment	Primary key	17302
GO_ori_id	int(11) unsigned	NOT NULL default '0'	Original GO_id when a genomic object is updated	3475
S_id	int(11) unsigned	default NULL	Foreign key (Sequence table primary key)	2
A_id	int(11) unsigned	default NULL	Foreign key (Annotator table primary key)	2
CA_id	int(11) unsigned	default NULL	Foreign key (Compare_Annotation table primary key)	0
GO_type	varchar(20)	default NULL	Type CDS, fCDS, cCDS, rRNA, tRNA	CDS
GO_frame	enum('+1"+2"+3"-1"-2"-3')	default NULL	Frame for the CDS and strand for the other objects (+1 or -1)	-2
GO_begin	int(11) unsigned	default NULL	Begin position	3400322
GO_end	int(11) unsigned	default NULL	End position	3402280
GO_annot_status	enum('finished"inProgress')	default NULL	The annotateur need to know the status of annotation (or re-annotation)	Curated
GO_mutation	enum('frameshift"pseudo"partial"no')	default NULL	Authentic frameshift is rare occurrence (e.g. *prfB* in *B. subtilis* should have a 'frameshift' status). More often, pseudogene is observed (e.g. *argD* in *Y. pestis*). There are several type of frameshifts (see the table ProFED). Another type of mutation (deletion) create a 'partial' CDS (e.g. *glpD* in *Y. pestis*). Finally, a point mutation (or stop in frame) is declared as 'pseudo'.	no
GO_label	varchar(30)	default NULL	Usually catenate a species code (or label_tag) and a formated number (with 0 upstream)	ACIAD3475
GO_gene_name	varchar(30)	default NULL	Gene Symbol	acs
GO_synonyms	varchar(255)	default NULL	Gene Symbol synonyms	acsA
GO_product	text	default NULL	Gene product name	acetyl-CoA synthetase
GO_product_type	text	default NULL	Carrier' ('c'), 'enzyme' ('e'), 'factor' ('f'), 'extrachromosomal origin' ('h'), etc. (see Product_Type table)	e : enzyme
GO_note	text	default NULL	Generally give the best similarity	similar to Acetyl-CoA synthase from Vibrio cholerae, swall : Q9KV59 (666 aa), Evalue = 2.76159e-257, %identity = 65.31 …
GO_comments	text	default NULL	Annotator comments	The E.coli enzyme is authentified.
GO_function	text	default NULL	Functional classification of M. Riley (see the Classification_Riley table)	1.1.1 : carbon compounds ; 1.5.4 : fatty acid and phosphatidic acid ; 1.3.4 : tricarboxylic acid cycle ;
GO_localization	text	default NULL	Unknown' (1), 'Cytoplasmic' (2), Inner membrane protein (5), etc. (see Localization table)	2 : Cytoplasmic
GO_class	text	default NULL	Doubtful CDS' (6), 'Gene remnant' (7) (see Status_Class table)	2a : Function of homologous gene experimentally demonstrated in an other organism
GO_length	int(11) unsigned	default NULL	Genomic object length in base pairs	1959
GO_ec	varchar(255)	default NULL	Enzyme Commission number	6.2.1.1
GO_evidence	enum('automatic_prediction"manual_annotation"experimental_validation')	default NULL	The status 'automatic_prediction' is for a raw biocomuting prediction. 'manual_annotation' is used when a validation in silico is done by a bioanalyst expert. 'experimental_validation' is used when a validation in vitro or in vivo is made by a biologist expert.	experimental_validation
GO_update	enum('current"obsolete')	default NULL	For exemple, 'Obsolete' update could be linked with 'automatic' evidence whereas 'current' update could be linked with 'validated' evidence	current
GO_creation_date	datetime	default NULL	Genomic object creation date	26/09/2003 14:19
GO_CA_status	enum('BANKnoCA"AGCnoCA"COMMON"uniqBANK"wrongBANK"suspiciousBANK"uniqAGC"newAGC"ambiguousAGC')	default NULL	To know where the genomic object is comming from (source) and the status of re-annotation if the annotation comparaison has been made	AGCnoCA

TAB. F.1 – Description détaillée des champs de la table Genomic_object

Annexe G

AMIGene

G.1 Algorithme des heuristiques

Choix des valeurs des paramètres de la stratégie *AMIGene* (Fig. G.1 p. 414) :

1. La longueur des CDS, L, est de 63 pb *sure_L* (codon de terminaison inclus et codon d'initiation le plus en 5′ ou *Leftmost Start* (LS)).
2. Le nombre de matrices d'occurrences d'oligonucléotides, c, dépend du nombre de classe de CDS homogènes dans leur usage des codons synonymes
3. Pour le réajustement du codon d'initiation ou *AMIGene Start* (AS), il y a cinq paramètres : *climb_P* (Pc pour repérer la position correspondant à la montée de la courbe de probabilités de codage), *diff_Pc* (pour réajuster le codon d'initiation la différence entre asPc et lsPc doit être supérieure à ce seuil), *sure_L*, *sure_Pc* et *prob_L*.
4. Le seuil de probabilité *sure_Pc* a varié entre 0.35 et 0.9. La probabilité moyenne de codage d'une CDS est « forte » quand elle est supérieure à ce seuil. Ce seuil permet de sélectionner définitivement les CDS sans tenir compte de leur environnement avec les autres CDS (mis à part les problèmes d'inclusion). En général une CDS dont la probabilité moyenne de codage est supérieure à 0.5 est une CDS avérée.
5. Le seuil de probabilité *prob_Pc* a varié entre 0.15 et 0.47. La probabilité moyenne de codage d'une CDS est « moyenne » quand elle est comprise entre ce seuil et le seuil de probabilité 1. Elle est « faible » quand elle est inférieure à ce seuil. Cet ordre de grandeur a été choisi dans le but de rechercher des erreurs d'annotations évidentes sur des génomes déjà publiés (le groupe de gènes avérés pour la phase d'apprentissage de la méthode GM est largement suffisant). Cependant selon les génomes, les plateaux GM sont plus ou moins marqués, ce qui explique la variation de la valeur de 0.15 à 0.3. Dans le cadre d'une première annotation sur un génome qui vient d'être assemblé, cette valeur peut descendre jusqu'à 0.005 ; ce seuil élimine alors les CDS dont la probabilité moyenne est nulle. Le seuil de longueur *prob_L* a varié de 114 à 213. La longueur d'une CDS '*probable*' doit être supérieure à ce seuil.

```
1. {longueur des CDS, L (1)}
Exécution de Prokov_Orf
2. {nombre de matrices d'occurrences d'oligonucléotides, c (2)}
Exécution de c Prokov_Curve
3. {combinaison des résultats des étapes 1 et 2: probabilité moyenne de codage des CDS, Pc}
Calcul de Pc et réajustement de la position du codon d'initiation (c AMI_CDS_score) (3)
4. {combinaison des c résultats de l'étape 3}
Conservation de la meilleure Pc et du numéro de matrice correspondant (AMI_CDS_combine)
5. {filtrage des CDS en fonction de L et Pc}
Constitution de deux listes de CDS triées par début (AMI_L_Pc_filter):
l_CDS_sures rassemble les CDS_sures ayant une Pc " forte " (4)
l_CDS_probables rassemble les CDS_prob ayant une Pc et une L " moyennes " (4) (5)
Au passage, élimination des CDS ayant une probabilité de codage " faible " (5)
6. {élimination des CDS_sures qui sont incluses dans d'autres CDS_sures}
AMI_sure_I_filter:
 Pour chaque CDS_courante de l_CDS_sures Faire
  Pour chaque CDS_incluse de l_CDS_sures incluse dans CDS_courante Faire
   Si CDS_incluse et CDS_courante sont en sens contraire
    Alors élimination de CDS_incluse
    Sinon Si le % d'inclusion de la CDS_courante est trop faible (6) (7)
           Alors élimination de CDS_incluse
          Fsi
   Fsi
  Fait
 Fait
7. {élimination des CDS_prob qui recouvrent des CDS_sures (inclusions ou chevauchements)}
AMI_sure_prob_IO_filter:
 Pour chaque CDS_prob de l_CDS_probables Faire
  Pour chaque CDS_sure de l_CDS_sures  telle que CDS_sure recouvre CDS_prob Faire
   Cas  a. CDS_prob incluse dans CDS_sure : élimination de CDS_prob
        b. CDS_sure et CDS_prob se chevauchent dans le sens contraire :
           Si le % de chevauchement de la CDS_prob est trop élevé (8)
            Alors élimination de CDS_prob
           Fsi
        Sinon {CDS_sure incluse dans CDS_prob ou elles se chevauchent dans le même sens} : rien
   Fcas
  Fait
 Fait
8. {élimination des CDS_prob qui ont des recouvrements importants avec d'autres CDS_prob, dans
    les régions où il y a en majorité des CDS_prob mais sans générer de trou d'annotation}
AMI_prob_glob_IO_filter:
 Pour chaque CDS_courante de l_CDS_probables Faire
  Pour chaque CDS_recouverte de l_CDS_probables telle que CDS_recouverte
  recouvre CDS_courante Faire
   Calcul du score total de recouvrement de la CDS_courante
  Fait
 Fait
 Pour chaque CDS_courante de l_CDS_probables Faire
  Si son score total de recouvrement est trop élevé (9)
   Alors Si la CDS_suivante a un score trop élevé et chevauche la CDS_courante
          Alors CDS_choix = CDS_courante
                Tant que CDS_suivante a un score trop élevé et chevauche la CDS_courante Faire
                 Si la Pc de la CDS_choix est inférieure à la Pc de la CDS_suivante
                  Alors élimination de la CDS_choix et CDS_choix = CDS_suivante
                  Sinon élimination de la CDS_suivante
                 Fsi
                Fait
          Sinon élimination de la CDS_courante
         Fsi
  Fsi
 Fait
```

FIG. G.1 – Algorithme d'*AMIGene*

6. Le seuil de pourcentage *sure_ss_I* a varié de 3 à 21 %. Pourcentage de recouvrement de la CDS courante avec une autre CDS = (taille du recouvrement entre les deux CDS / taille de la CDS courante) * 100 Valeur déterminée après études des « vrais » frameshifts de Bacillus subtilis. 10 % c'est le pourcentage d'inclusion du plus petit décalage du cadre de lecture compensé de *B. subtilis*. Pour qu'une CDS '*sure*' incluse dans une CDS '*sure*' dans le même sens soit conservée, le pourcentage d'inclusion doit supérieur ou égal à *sure_ss_I*.
7. Le seuil de pourcentage *sure_os_I* a varié de 30 à 56 %. Pour qu'une CDS '*sure*' incluse dans une CDS '*sure*' dans le sens contraire soit conservée, le pourcentage d'inclusion doit supérieur ou égal à *sure_os_I*. Ce cas de figure met généralement en évidence des artefacts mais il est intéressant car il pointe sur des régions de basse complexité compositionnelle et de répétitions de courte périodicité générant des ombres du codant (d'où l'inclusion en sens contraire).
8. Le seuil de pourcentage *sure_prob_O* a varié de 0 à 37%. Le pourcentage de chevauchement autorisé entre une CDS '*probable*' et une CDS '*sure*' en sens inverse a été ajusté *in silico*. Le seuil de pourcentage de chevauchement entre une CDS '*probable*' et une CDS '*sure*' dans le même sens a été retiré pour autoriser les décalages du cadre de lecture (on tolère les chevauchements quand les CDS sont dans le même sens).
9. Le seuil de score *prob_glob_IO* a varié de 49 à 103. Le score total de recouvrement de la CDS courante c'est la somme de tous ses scores de recouvrement (quelque soit l'orientation des recouvrements même sens ou sens contraire). On permet qu'une CDS probable ait un recouvrement total qui représente au maximum 1 fois sa taille (dans le cas contraire elle est éliminée).

D'un génome à un autre les seuls paramètres qui ont variés sont les seuils de probabilités 1 et 2. Le seuil de probabilité 2 doit être compris entre 0.15 et 0.3. Le seuil de probabilité 1 est égal au seuil de probabilité 2 plus 0.2. Les valeurs des seuils de probabilité doivent être maximales mais le nombre de CDS sélectionnées multiplié par 1000 pb doit couvrir au minimum 90% de la taille du génome.

G.2 Optimisation des paramètres de la phase de reconnaissance des paramètres

Les résultats de l'optimisation automatique des paramètres de *compute_Pc* chez *B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv sont présentés pour R2 et R10 (Tab. G.1 p. 416). Pour *B. subtilis* et *E. coli* K-12, les valeurs originales des quatre paramètres ont servies à l'initialisation pour R5. Une seconde itération permet de vérifier que l'équilibre est atteint. Pour R2 et R10, nous testons deux conditions initiales, le jeu original (en jaune foncé) et le jeu optimal pour R5, pour lesquelles nous ne faisons qu'une itération (*e.g.* les valeurs de paramètre optimales pour *E. coli* K-12 R5 : *climb_P* = 0,965, *diff_Pc* = 0,07, *sure_Pc* = 0,7 et *prob_L* = 195 pb).
Pour *M. tuberculosis* H37Rv, nous avons recommencé l'optimisation avec quatre matrices en prenant

comme valeurs initiales, les valeurs optimales obtenues pour R10 avec trois matrices (nous effectuons deux itérations).

BACSU p=2 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,950	**0,090**	**0,7**	**150**	**98,44**	**92,01**	**90,69**	**3,703**		
0,965	0,090	0,7	150	98,44	92,05	90,73	3,689	0,014	
0,965	**0,140**	0,7	150	98,39	92,17	90,81	3,681	0,008	
0,965	0,070	**0,35**	150	98,37	92,55	91,15	3,571	0,110	
0,965	**0,070**	**0,9**	**123**	**98,47**	**92,45**	**91,14**	**3,539**	0,032	4,43
0,965	0,070	0,9	123	98,47	92,45	91,14	3,539	0,000	
0,965	**0,070**	0,9	123	98,47	92,45	91,14	3,539	0,000	
0,965	0,070	**0,9**	123	98,47	92,45	91,14	3,539	0,000	
0,965	**0,070**	**0,9**	**123**	**98,47**	**92,45**	**91,14**	**3,539**	0,000	0,00

BACSU p=10 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,950	**0,090**	**0,7**	**150**	**98,44**	**92,01**	**90,69**	**2,144**		
0,965	0,090	0,7	150	98,44	92,05	90,73	2,140	0,004	
0,965	**0,030**	0,7	150	98,49	91,78	90,51	2,120	0,020	
0,965	0,070	**0,9**	150	98,47	92,07	93,77	2,116	0,004	
0,965	**0,070**	**0,9**	**123**	**98,47**	**92,45**	**91,14**	**2,082**	0,034	2,90
0,965	0,070	0,9	123	98,47	92,45	91,14	2,082	0,000	
0,965	**0,070**	0,9	123	98,47	92,45	91,14	2,082	0,000	
0,965	0,070	**0,9**	123	98,47	92,45	91,14	2,082	0,000	
0,965	**0,070**	**0,9**	**123**	**98,47**	**92,45**	**91,14**	**2,082**	0,000	0,00

ECOLI p=2 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,950	**0,090**	**0,7**	**150**	**98,63**	**91,41**	**90,26**	**3,780**		
0,965	0,090	0,7	150	98,65	91,43	90,30	3,757	0,023	
0,965	**0,220**	0,7	150	98,55	91,91	90,17	3,661	0,096	
0,965	0,070	**0,85**	150	98,65	91,41	90,17	3,763	-0,103	
0,965	**0,070**	**0,7**	**195**	**98,70**	**91,54**	**90,44**	**3,690**	0,073	2,37
0,965	0,070	0,7	195	98,70	91,54	90,44	3,690	0,000	
0,965	**0,220**	0,7	195	98,55	91,93	90,71	3,654	0,036	
0,965	0,220	**0,75**	195	98,67	91,66	90,54	3,666	-0,012	
0,965	**0,220**	**0,75**	**195**	**98,67**	**91,66**	**90,54**	**3,666**	0,000	0,66

ECOLI p=10 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,950	**0,090**	**0,7**	**150**	**98,63**	**91,41**	**90,26**	**2,031**		
0,965	0,090	0,7	150	98,65	91,43	90,30	2,007	0,024	
0,965	**0,010**	0,7	150	98,74	91,02	89,98	1,958	0,049	
0,965	0,070	**0,7**	150	98,70	91,26	90,17	1,980	-0,022	
0,965	**0,070**	**0,7**	**195**	**98,70**	**91,54**	**90,44**	**1,955**	0,026	3,75
0,965	0,070	0,7	195	98,70	91,54	90,44	1,955	0,000	
0,965	**0,070**	0,7	195	98,70	91,54	90,44	1,955	0,000	
0,965	0,070	**0,7**	195	98,70	91,54	90,44	1,955	0,000	
0,965	**0,070**	**0,7**	**195**	**98,70**	**91,54**	**90,44**	**1,955**	0,000	0,00

MYCTU p=2 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,915	**0,050**	**0,70**	**150**	**98,02**	**91,18**	**89,53**	**4,259**		
0,950	0,050	0,70	150	98,02	91,26	89,61	4,231	0,028	
0,950	**0,160**	0,70	150	97,75	92,36	90,44	4,049	0,182	
0,950	0,160	**0,75**	150	97,75	92,38	90,46	4,041	0,007	
0,950	**0,160**	**0,75**	**150**	**97,75**	**92,38**	**90,46**	**4,041**	0,000	5,118
0,965	0,160	0,75	150	97,77	92,41	90,50	4,017	0,025	
0,965	**0,160**	0,75	150	97,77	92,41	90,50	4,017	0,000	
0,965	0,160	**0,75**	150	97,77	92,41	90,50	4,017	0,000	
0,965	**0,160**	**0,75**	**150**	**97,77**	**92,41**	**90,50**	**4,017**	**0,000**	0,610

MYCTU p=10 L

Climb-P	Diff-Pc	Sure	Prob	Sn	Sp	Pr	R	M	D
0,915	**0,050**	**0,70**	**150**	**98,02**	**91,18**	**89,53**	**2,599**		
0,940	0,050	0,70	150	98,05	91,16	89,53	2,578	0,021	
0,940	**0,050**	0,70	150	98,05	91,16	89,53	2,578	0,000	
0,940	0,050	**0,75**	150	98,05	91,18	89,55	2,577	0,002	
0,940	**0,050**	**0,75**	**147**	**98,05**	**91,20**	**89,57**	**2,575**	0,002	0,955
0,940	0,050	0,75	147	98,05	91,20	89,57	2,575	0,000	
0,940	**0,050**	0,75	147	98,05	91,20	89,57	2,575	0,000	
0,940	0,050	**0,75**	147	98,05	91,20	89,57	2,575	0,000	
0,940	**0,050**	**0,75**	**147**	**98,05**	**91,20**	**89,57**	**2,575**	0,000	0,000

TAB. G.1 – Optimisation des paramètres de *compute_Pc*

G.3 Optimisation des paramètres de la phase de filtrage des paramètres

Les résultats de l'optimisation automatique des paramètres d'*AMI_filter_CDS* chez *B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv sont présentés pour $p = 2$ et $p = 10$.

Pour *B. subtilis*, les valeurs initiales ont été choisies d'après l'analyse des caractéristiques générales des jeux de CDS totales et de référence (TAB. G.2 p. 417 et voir p. 251). Pour $p = 2$, il n'y a pas d'amélioration (minimisation de R2) durant la troisième itération. Des valeurs de paramètres très sévères sont *prob_Pc* = 0,41 *sure_Pc* = 0,68 *prob_L* =165 *sure_ss_I* = 1, *sure_os_I* =1 *sure_prob_os_O* = 0 et *prob_glob_IO* =0,64. Des valeurs de paramètres sévères sont *prob_Pc* = 0,41 *sure_Pc* =0,68 *prob_L* =165, *sure_ss_I* = 0 *sure_os_I* = 0,2 *sure_prob_os_O* = 0 et *prob_glob_IO* = 0,65. Toutes les variations entre ces minimums et ces maximums donnent le même jeu de CDS prédites par *AMI_filter_CDS*. Pour $p = 10$, il n'y a pas d'amélioration durant la seconde itération.

BACSU	p = 2	0,965	0,07	0,9	123							
prob-Pc	**sure-Pc**	**prob-L**	**sure-ss-l**	**sure-os-l**	**sure-prob-os-O**	**prob-glob-lO**	**Sn**	**Sp**	**Pr**	**R2**	**M**	**D**
0,4	**0,7**	**150**	**0,1**	**0,3**	**0,2**	**0,7**	**98,08**	**93,76**	**92,07**	**3,363**		
0,41	0,7	150	0,1	0,3	0,2	0,7	98,00	93,95	92,19	3,348	0,015	
0,41	**0,67**	150	0,1	0,3	0,2	0,7	98,15	93,74	92,11	3,320	0,028	
0,41	0,67	**165**	0,1	0,3	0,2	0,7	97,91	94,72	92,84	3,156	0,164	
0,41	0,67	165	**1**	0,3	0,2	0,7	97,91	94,72	92,84	3,156	0,000	
0,41	0,67	165	1	**1**	0,2	0,7	97,91	94,72	92,84	3,156	0,000	
0,41	0,67	165	1	1	**0**	0,7	97,86	95,01	93,07	3,093	0,063	
0,41	**0,67**	**165**	**1**	**1**	**0**	**0,64**	**97,86**	**95,03**	**93,09**	**3,085**	0,008	8,26
0,41	0,67	165	1	1	0	0,64	97,86	95,03	93,09	3,085	0,000	
0,41	**0,68**	165	1	1	0	0,64	97,76	95,23	93,20	3,084	0,001	
0,41	0,68	**165**	1	1	0	0,64	97,76	95,23	93,20	3,084	0,000	
0,41	0,68	165	**0**	1	0	0,64	97,76	95,23	93,20	3,084	0,000	
0,41	0,68	165	0	**0,2**	0	0,64	97,76	95,23	93,20	3,084	0,000	
0,41	0,68	165	0	0,2	**0**	0,64	97,76	95,23	93,20	3,084	0,000	
0,41	**0,68**	**165**	**0,05**	**0,3**	**0**	**0,65**	**97,76**	**95,23**	**93,20**	**3,084**	0,000	0,04
BACSU	**p = 10**	0,965	0,07	0,9	123							
prob-Pc	**sure-Pc**	**prob-L**	**sure-ss-l**	**sure-os-l**	**sure-prob-os-O**	**prob-glob-lO**	**Sn**	**Sp**	**Pr**	**R10**	**M**	**D**
0,3	**0,6**	**120**	**0,1**	**0,3**	**0,15**	**0,9**	**99,03**	**86,88**	**86,14**	**2,078**		
0,35	0,6	150	0,1	0,3	0,15	0,8	98,93	88,78	87,94	1,994	0,084	
0,35	**0,67**	150	0,1	0,3	0,15	0,8	98,95	89,14	88,30	1,940	0,054	
0,35	0,67	**114**	0,1	0,3	0,15	0,8	99,03	88,76	87,99	1,908	0,032	
0,35	0,67	114	**1**	0,3	0,15	0,8	99,03	88,76	87,99	1,908	0,000	
0,35	0,67	114	1	**1**	0,15	0,8	99,03	88,76	87,99	1,908	0,000	
0,35	0,67	114	1	1	**0,05**	0,8	99,03	89,07	88,29	1,879	0,028	
0,35	**0,67**	**114**	**0,05**	**0,3**	**0,05**	**0,86**	**99,03**	**89,13**	**88,35**	**1,874**	0,005	9,83

TAB. G.2 – Optimisation des paramètres d'*filter_L_Pc* chez *B. subtilis*

Abréviations : p, facteur de pondération ; Sn, pourcentage de sensibilité d'*AMI_filter_CDS* ; Sp pourcentage de spécificité ; Rp, fonction de risque à minimiser pour trouver les valeurs optimales, M, minimisation (différence entre deux Risques successifs) ; D, critére d'arrêt ($((R_i - R_{i+1}) * 100)/R_i \leq 0,1$), CDSa pour les CDS d'*AMIGene* et CDSb pour les CDS de référence.

Pour *E. coli* K-12, les valeurs initiales du jeu de paramètres d'*AMI_filter_CDS* sont les valeurs originales de l'*Article III* p. 289 sauf que l'on a augmenté *prob_Pc* et *sure_Pc* et que l'on est passé de *sure_prob_O* à *sure_prob_os_O* car depuis ce travail de réannotation, nous avons mis en place quelques améliorations (Tab. G.3 p. 419 et voir p. 277 et p. 268 ; *prob_Pc* = 0,4, *sure_Pc* = 0,7, *prob_L* = 150 pb, *sure_ss_I* = 0,1, *sure_os_I* = 0,3, *sure_prob_os_O* = 0,5, *prob_glob_IO* = 0,8). Des valeurs de paramètres très sévères pour $p = 2$ dans les conditions initiales originales sont *prob_Pc* =0,47 *sure_Pc* =0,7 *prob_L* =189 *sure_ss_I* =0,03 *sure_os_I* =0,7 *sure_prob_os_O* =0,16 *prob_glob_IO* =0,46 ; tandis que des valeurs sévères : *prob_Pc* =0,47 *sure_Pc* =0,7 *prob_L* =189 *sure_ss_I* =0 *sure_os_I* =0,29 *sure_prob_os_O* =0,16 *prob_glob_IO* =0,6. Des valeurs de paramètres sévères pour $p = 10$ dans les conditions initiales originales sont *prob_Pc* = 0,41 *sure_Pc* = 0,68 *prob_L* =165 *sure_ss_I* = 1, *sure_os_I* =1 *sure_prob_os_O* = 0 et *prob_glob_IO* =0,64 ; tandis que des valeurs peu sévères sont *prob_Pc* = 0,41 *sure_Pc* =0,68 *prob_L* =165, *sure_ss_I* = 0 *sure_os_I* = 0,2 *sure_prob_os_O* = 0 et *prob_glob_IO* = 0,65.

Pour *M. tuberculosis* H37Rv, les valeurs initiales ont été choisies d'après l'analyse des caractéristiques générales des jeux de CDS totales et de référence (Tab. G.4 p. 420). Des valeurs de paramètres très sévères pour $p = 2$ sont *prob_Pc* =0,36 *sure_Pc* =0,66 *prob_L* =225 *sure_ss_I* =1 *sure_os_I* =1 *sure_prob_os_O* =0,16 *prob_glob_IO* =0,75 ; tandis que des valeurs sévères sont *prob_Pc* =0,36 *sure_Pc* =0,66 *prob_L* =225 *sure_ss_I* =0,14 *sure_os_I* =0,31 *sure_prob_os_O* =0,17 *prob_glob_IO* =0,81. Il n'y a pas d'amélioration durant la troisième itération. Des valeurs de paramètres sévères pour $p = 10$ sont *prob_Pc* =0,27 *sure_Pc* =0,57 *prob_L* =165 *sure_ss_I* =1 *sure_os_I* =1 *sure_prob_os_O* =0,14 *prob_glob_IO* =0,5 ; tandis que des valeurs peu sévères sont *prob_Pc* =0,27 *sure_Pc* =0,57 *prob_L* =165 *sure_ss_I* =0,2 *sure_os_I* =0,54 *sure_prob_os_O* =0,14 *prob_glob_IO* =0,54. Il n'y a pas d'amélioration durant la deuxième et la troisième itération.

ECOLI	p = 2	0,965	0,22	0,75	195							
prob-Pc	sure-Pc	prob-L	sure-ss-l	sure-os-l	sure-prob-os-O	prob-glob-IC	Sn	Sp	Pr	R2	M	D
0,4	**0,7**	**150**	**0,1**	**0,3**	**0,5**	**0,8**	**98,63**	**92,36**	**91,19**	**3,462**		
0,53	0,7	150	0,1	0,3	0,5	0,8	97,68	95,99	93,85	2,884	0,578	
0,53	**0,7**	150	0,1	0,3	0,5	0,8	97,68	95,99	93,85	2,884	0,000	
0,53	0,7	**189**	0,1	0,3	0,5	0,8	97,37	96,61	94,16	2,883	0,001	
0,53	0,7	189	**0,03**	0,3	0,5	0,8	97,39	96,59	94,16	2,874	0,009	
0,53	0,7	189	0,03	**0,7**	0,5	0,8	97,39	96,59	94,16	2,874	0,000	
0,53	0,7	189	0,03	0,7	**0,16**	0,8	97,37	96,80	94,33	2,822	0,052	
0,53	**0,7**	**189**	**0,03**	**0,7**	**0,16**	**0,46**	**97,37**	**96,84**	**94,37**	**2,807**	0,015	18,92
0,47	0,7	189	0,03	0,7	0,16	0,46	97,77	96,33	94,33	2,708	0,099	
0,47	**0,7**	189	0,03	0,7	0,16	0,46	97,77	96,33	94,33	2,708	0,000	
0,47	0,7	**189**	0,03	0,7	0,16	0,46	97,77	96,33	94,33	2,708	0,000	
0,47	0,7	189	**0**	0,7	0,16	0,46	97,77	96,33	94,33	2,708	0,000	
0,47	0,7	189	0	**0,29**	0,16	0,46	97,77	96,33	94,33	2,708	0,000	
0,47	0,7	189	0	0,29	**0,16**	0,46	97,77	96,33	94,33	2,708	0,000	
0,47	**0,7**	**189**	**0,03**	**0,3**	**0,16**	**0,6**	**97,77**	**96,33**	**94,33**	**2,708**	0,000	3,54
ECOLI	p = 10	0,965	0,07	0,7	195							
prob-Pc	sure-Pc	prob-L	sure-ss-l	sure-os-l	sure-prob-os-O	prob-glob-IC	Sn	Sp	Pr	R10	M	D
0,4	**0,7**	**150**	**0,1**	**0,3**	**0,5**	**0,8**	**98,77**	**91,87**	**90,82**	**1,860**		
0,4	0,7	150	0,1	0,3	0,5	0,8	98,77	91,87	90,82	1,860	0,000	
0,4	**0,62**	150	0,1	0,3	0,5	0,8	98,84	91,51	90,53	1,827	0,033	
0,4	0,62	**153**	0,1	0,3	0,5	0,8	98,84	91,69	90,71	1,811	0,016	
0,4	0,62	153	**0,2**	0,3	0,5	0,8	98,84	91,69	90,71	1,811	0,000	
0,4	0,62	153	0,2	**1**	0,5	0,8	98,84	91,71	90,73	1,809	0,002	
0,4	0,62	153	0,2	1	**0**	0,8	98,72	93,26	92,15	1,776	0,033	
0,4	**0,62**	**153**	**0,2**	**1**	**0**	**0,74**	**98,72**	**93,26**	**92,15**	**1,776**	0,000	4,52
0,4	0,62	153	0,2	1	0	0,74	98,72	93,27	92,21	1,776	0,000	
0,4	**0,62**	153	0,2	1	0	0,74	98,72	93,27	92,21	1,776	0,000	
0,4	0,62	**141**	0,2	1	0	0,75	98,82	92,38	91,37	1,770	0,006	
0,4	0,62	141	**0,04**	1	0	0,75	98,82	92,38	91,37	1,770	0,000	
0,4	0,62	141	0,04	**0,56**	0	0,75	98,82	92,38	91,37	1,770	0,000	
0,4	0,62	141	0,04	0,56	**0**	0,75	98,82	92,38	91,37	1,770	0,000	
0,4	**0,62**	**141**	**0,05**	**0,56**	**0**	**0,75**	**98,82**	**92,38**	**91,37**	**1,770**	0,000	0,35

TAB. G.3 – Optimisation des paramètres d'*filter_L_Pc* chez *E. coli* K-12

Pour les abréviations voir la légende du tableau G.2 p. 417.

MYCTU	p = 2	0,965	0,16	0,75	150							
prob-Pc	sure-Pc	prob-L	sure-ss-I	sure-os-I	sure-prob-os-O	prob-glob-IO	Sn	Sp	Pr	R2	M	D
0,2	**0,65**	**192**	**0,1**	**0,3**	**0,2**	**0,7**	**97,70**	**89,50**	**87,65**	**5,036**		
0,36	0,65	192	0,1	0,3	0,2	0,7	97,32	94,37	91,98	3,662	1,374	
0,36	**0,67**	192	0,1	0,3	0,2	0,7	97,30	94,48	92,07	3,641	0,021	
0,36	0,67	**225**	0,1	0,3	0,2	0,7	96,90	95,79	92,94	3,472	0,170	
0,36	0,67	225	**0,05**	0,3	0,2	0,7	96,92	95,75	92,92	3,470	0,001	
0,36	0,67	225	0,05	**1**	0,2	0,7	96,92	95,75	92,92	3,470	0,000	
0,36	0,67	225	0,05	1	**0,16**	0,7	96,90	95,86	93,01	3,448	0,022	
0,36	**0,67**	**225**	**0,05**	**1**	**0,16**	**0,75**	**96,92**	**95,87**	**93,03**	**3,431**	0,017	31,87
0,36	0,67	225	0,05	1	0,16	0,75	96,92	95,87	93,03	3,431	0,000	
0,36	**0,66**	225	0,05	1	0,16	0,75	96,95	95,82	93,01	3,430	0,001	
0,36	0,66	**225**	0,05	1	0,16	0,75	96,95	95,82	93,01	3,430	0,000	
0,36	0,66	225	**0,14**	1	0,16	0,75	96,92	95,89	93,06	3,423	0,007	
0,36	0,66	225	0,14	**0,31**	0,16	0,75	96,92	95,89	93,06	3,423	0,000	
0,36	0,66	225	0,14	0,31	**0,17**	0,75	96,92	95,89	93,06	3,423	0,000	
0,36	**0,66**	**225**	**0,14**	**0,31**	**0,16**	**0,8**	**96,92**	**95,89**	**93,06**	**3,423**	0,000	0,23
MYCTU	p = 10	0,94	0,05	0,75	147							
prob-Pc	sure-Pc	prob-L	sure-ss-I	sure-os-I	sure-prob-os-O	prob-glob-IO	Sn	Sp	Pr	R10	M	D
0,05	**0,55**	**162**	**0,1**	**0,3**	**0,15**	**0,9**	**98,72**	**73,51**	**72,82**	**3,569**		
0,27	0,6	162	0,1	0,3	0,15	0,9	98,42	88,00	86,78	2,524	1,045	
0,27	**0,57**	162	0,1	0,3	0,15	0,9	98,45	88,46	87,25	2,460	0,064	
0,27	0,57	**165**	0,1	0,3	0,15	0,9	98,42	88,78	87,53	2,454	0,006	
0,27	0,57	165	**1**	0,3	0,15	0,9	98,42	88,82	87,57	2,450	0,004	
0,27	0,57	165	1	**1**	0,15	0,9	98,42	88,88	87,63	2,445	0,005	
0,27	0,57	165	1	1	**0,14**	0,9	98,42	88,94	87,69	2,439	0,005	
0,27	**0,57**	**165**	**0,2**	**0,54**	**0,14**	**0,54**	**98,42**	**89,32**	**88,06**	**2,404**	0,035	32,63

TAB. G.4 – Optimisation des paramètres d'*filter_L_Pc* chez *M. tuberculosis* H37Rv
Pour les abréviations voir la légende du tableau G.2 p. 417.

	prob-Pc	sure-Pc	prob-L	sure-ss-I	sure-os-I	sure-prob-os-O	prob-glob-IO	Sn	Sp	Pr	R	Nb CDSa
ECOLI R2	0,47	0,75	165	0,03	0,3	0,15	0,49	97,99	95,92	94,06	2,705	4310
ECOLI R10	0,4	0,7	138	0,05	0,3	0,02	0,75	98,84	92,34	91,35	1,752	4516
MYCTU R1(	0,3	0,51	204	0,24	0,56	0,5	1,01	98,40	89,54	88,26	2,407	4390
	prob-Pc	**sure-Pc**	**prob-L**	**sure-ss-I**	**sure-os-I**	**sure-prob-os-O**	**prob-glob-IO**	**Sn**	**Sp**	**Pr**	**R**	**Nb CDSa**
ECO57 R2	0,47	0,75	165	0,03	0,3	0,15	0,49	95,36	96,83	92,47	4,153	5261
ECO57 R10	0,4	0,7	138	0,05	0,3	0,02	0,75	96,76	93,14	90,32	3,568	5550
MYCTC R1(	0,3	0,51	204	0,24	0,56	0,5	1,01	92,80	90,24	84,33	7,436	4426

TAB. G.5 – Autres résultats de validation automatique d'*AMI_filter_CDS*
A) Des valeurs finales de l'optimisation des paramètres d'*AMI_filter_CDS* chez *B. subtilis*, *E. coli* K-12 et *M. tuberculosis* H37Rv.
B) Des résultats de la validation automatique des paramètres d'*AMI_filter_CDS* chez *B. halodurans*, *E. coli* O157:H7 EDL933 et *M. tuberculosis* CDC1551.
Pour les abréviations voir la légende du tableau G.2 p. 417.

Annexe H

Tests statistiques non paramétriques

Les paramètres des distributions de variables dans des populations doivent être connus lorque l'on cherche à comparer ces distributions au moyen de tests statistiques paramétriques. En revanche, les méthodes non paramétriques ne reposent pas sur l'estimation de paramètres tels que la moyenne ou l'écart type décrivant la distribution de la variable étudiée dans la population (méthodes libre de paramètres ou de distributions).

H.1 Test U de Mann - Whitney (u-test)

Il est utile dans les mêmes cas qu'un test de Student (t-test). C'est la version non paramétrique du t-test de deux groupes indépendants. Le t-test teste l'hypothèse que les moyennes de deux groupes sont différentes, les deux groupes suivant une loi normale. Dans l'exemple de deux distributions, l'une représentant un groupe de patients ayant reçu une drogue et l'autre ayant reçu un placebo, H0 est l'hypothèse neutre *les deux distribution sont similaires* et H1 est l'hypothèse dirigée *les deux distribution sont différentes*. Si la P-value d'un t-test est significative (P-value $<$ 0.05) alors on rejette H0 et accepte H1, les deux distributions sont donc différentes. Au contraire, Si la P-value d'un t-test n'est pas significative (P-value $>=$ 0.05) alors on accepte H0 et rejette H1, les deux distributions sont donc similaires. Le u-test teste que les distributions caractérisant les deux groupes sont différentes. Les restrictions de u-test sont que les deux groupes d'observations sont issus de distributions continues et sont indépendants à la fois à l'intérieur d'un groupe et entre les deux groupes. Puisque le u-test test ne compare pas les observations mais les rangs de ces observations il est donc résistant aux valeurs extrêmes dans l'un ou l'autre des deux groupes comparés.

Le t-test compare les moyennes alors que le u-test compare la somme des rangs. Le u-test est le plus puissant (ou sensible) des alternatives non paramétriques au t-test pour les échantillons indépendants. Il arrive que le u-test rejète l'hypothèse nulle (les distributions sont semblables) là ou le t-test l'aurait acceptée (les distributions sont différentes). Quand il y a plus de 20 observations, la distribution d'échantillonage de la Statistique U approche rapidement une distribution normale.

C'est pourquoi la statistique U (ajustée pour les exæquo) est accompagnée par valeur z (valeur d'une distribution normale) et de sa valeur p respective.

H.2 Test de Kolmogorov - Smirnov

Le KS teste si la distribution d'une variable continue est la même pour les deux groupes. Il teste donc l'hypothèse nulle que les distributions sont les mêmes à condition que les observations des distributions soient indépendantes l'une de l'autre. On compare les distributions au niveau de certains points et on considère donc la différence maximum entre les deux distributions. Ce ne sont pas les points des données qui sont comparés mais on calcule une fonction de ces points et ce sont ces fonctions qui sont comparées. Puisque ce test est basé sur la valeur maximale d'un jeu de nombres, il peut être largement influencé par les valeurs extrêmes et doit être utilisé avec précaution si on suspecte des valeurs extrêmes.

H.3 Test runs de Wald - Wolfowitz

Il teste si deux groupes d'observations ont été échantillonnés au hasard à partir de la même population. Ce test compare deux groupes qu'on pense être indépendant l'un de l'autre en combinant les données des deux groupes (seul le nombre de runs est important pas leur longueur). Si les deux échantillons viennent de distributions similaires (dans le texte de staview ils ont écrits différents mais je pense que c'est une erreur) on s'attend à avoir beaucoup de groupes de petits runs, tandis que si les observations d'un groupe ont tendance plus grande que celles de l'autre groupe on verra seulement un petit nombre de run dans les données. Puisque le test est basé sur les rangs il est résistant aux valeurs extrêmes. Le WW compare les données sur tout l'intervalle alors que le KS compare la différence maximale entre les distributions. S'il y a seulement une ou deux valeurs extrêmes le KS peut décider à tort que les deux distributions sont différentes.

Annexe I

Algorithmes pour l'attribution de statuts de réannotation

L'algorithme de l'attribution des statuts '*suspiciousBank*' ou '*wrongBank*' à certaines CDS uniques aux banques (GNF) de faible Pc, de courte longueur et sans similitude est décrit dans la figure I.1 p. 424. L'algorithme de l'attribution des status '*ambiguousAGC*' ou '*newAGC*' à certaines CDS uniques à *AMIGene* (NG) de forte Pc et de longueur importante est décrit dans la figure I.2 p. 425.

```
                 ------------------------------------------------------
                | Suspicious and Wrong Unique Bank CDS (swBK algorithm) |
                 ------------------------------------------------------

DATA = NGfile # Unique AMIGene CDS DATB = GNFfile # Unique Bank CDS
ua_Pc = 0,7
ub_Pc = 0,4       wrong_Pc = 0,1
ub_glob_IO = 0,5
suspi_L = 900     wrong_L = 162 # Program input parameters

#*****************************************************************************************
FUNCTION percentchev # compute covering percentage
{
 if(bA >= bB) {bC = bA} else{bC = bB}
 if(eA <= eB ) {eC = eA} else{eC = eB}
 return(((eC - bC) + 1) / lB)
}

#*****************************************************************************************
i = 0
while(DATA)
{
 if(Line is in the right format)
 {
  if(pA >= ua_Pc) # initialisation of variables for each NG (Pc>=0,7)
  {
   idA = field1 # CDSA identifiant
   bA = field2 # CDSA begin
   eA = field3 # CDSA end
   lA = field4 # CDSA length
   strA = field5 # CDSA strand
   fA = field6 # CDSA frame
   pA = field7 # CDSA coding probability
   mA = field8 # CDSA matrix
   simA = field9 # CDSA similarity
   evalA = field10 # CDSA best similarity Evalue
   nbsimA = field11 # CDSA similarity number
   store variables in TABA
   i++
  } # endif (pA >= ua_Pc)
 }else{print "Format error A"}
} # endwhile A

#*****************************************************************************************
nbA = i
while(DATB)
{
 if(Line is in the right format) #initialisation of variables for each GNF
 {
  idB bB eB lB strB fB pB mB simB evalB nbsimB # same as above for NG
  percenCT = 0 # total covering percentage
  idAT = "" # all idA covering current CDSB
  if((pB < ub_Pc) && (simB == "NO SIMILARITY FOUND") && (lB < suspi_L))
  {
   for(i=1 i<=nbA i++)
   {
    read TABA
    percentC = 0 status = ""
    # case1: B includes A or case2: B is included in A or
# case3: B overlaps A by the left or case4: B overlaps A by the right
    if(((bB <= bA) && (eB >= eA)) || ((bB > bA) && (eB < eA)) || ((bB < bA) && (eB > bA)
       && (eB < eA)) || ((eB > eA) && (bB > bA) && (bB < eA)))
    {
     percenCT = percenCT + percentchev(percentC)
     idAT = concat(idAT, " ", idA)
    } # endif covering
   } # endfor
   # if(percenCT > 0.5) wrong else depending on L and Pc wrong or suspi
   if((percenCT >= ub_glob_IO)||((pB < wrong_Pc)&&(lB < wrong_L))) {status = "WRONG"}
   else{status = "SUSPICIOUS"}
  } else{status= "NO STATUS"}
 } else{print "Format error B"} # endif modeleB
} # endwhile DATB
```

FIG. I.1 – Algorithme pour l'attribution de statuts 'faux' et 'douteux' à certaines CDS uniques aux annotations des banques.

```
DATA = NGfile # Unique AMIGene CDS DATB = GNFfile # Unique Bank CDS
ua_Pc = 0,7  ub_Pc = 0,4   ua_IO = 0,01  new_L = 300 # Program input parameters
FUNCTION percentchev # computes covering percentage
{if(bB >= bA){bC = bB} else{bC = bA} if(eB <= eA){eC = eB} else{eC = eA} return(((eC - bC) + 1)/lA)}
FUNCTION NAstatus # attribute CDSA status new or ambig depending on length or similarity
{if((lA >= new_L) || (simA == "SIMILARITY FOUND")){return("NEW")} else{return("AMBIGUOUS")}}
i = 0
while(DATB)
{
 if(Line is in the right format) # initialisation of variables for each GNF
 {
  idB = field1 # CDSB identifiant
  bB = field2 # CDSB begin
  eB = field3 # CDSB end
  lB = field4 # CDSB length
  strB = field5 # CDSB strand
  fB = field6 # CDSB frame
  pB = field7 # CDSB coding probability
  mB = field8 # CDSB matrix
  simB = field9 # CDSB similarity
  evalB = field10 # CDSB best similarity Evalue
  nbsimB = field11 # CDSB similarity number
  store variables in TABB
  i++
 } else{print "Format error B"}
} # endwhile B
nbB = i
while(DATA)
{
 if(Line is in the right format) # initialisation of variables for each NG
 {
  idA bA eA lA strA fA pA mA simA evalA nbsimA # same as above for GNF
  percenCT = 0 # total covering percentage
  statuT = "" # all status attributed to the current CDSA
  idBT = "" # concatenation of CDSB Identifiants covering current CDSA
  if(pA >= ua_Pc)
  {
   for(i=1 i<=nbB i++)
   {
    read TABB
    percentC = 0 status = ""
    # case1: A includes B or case2: A is included in B or
   # case3: A overlaps B by the left or case4: A overlaps B by the right
    if(((bA <= bB) && (eA >= eB)) || ((bA > bB) && (eA < eB)) || ((bA < bB) && (eA > bB)
       && (eA < eB)) || ((eA > eB) && (bA > bB) && (bA < eB)))
    {
     percenCT = percenCT + percentchev(percentC)
     idBT = concat(idBT, "", idB)
     if( pB < ub_Pc ) # PcB < 0,4
     {
      if( simB == "NO SIMILARITY FOUND") #sim status is "NOSIM"
       {statuT = concat(statuT, " ", NAstatus(status))}
      else # similarity status is " SIMTO "
      { # less than 1% of overlap
       if(percentC < ua_IO) {statuT = concat(statuT," ",NAstatus(status))}
       elsif(simA =="SIMILARITY FOUND"){statuT = concat(statuT," AMBIGUOUS")}
       else {statuT = concat(statuT, " NO STATUS")}
      }
     }
     else # PcB >= 0,4
     {
      if(percentC < ua_IO){statuT = concat(statuT," AMBIGUOUS")}
      else{statuT = concat(statuT, " NO STATUS")}
     } # endif (pB < ub_Pc)
    } # endif covering
   } # endfor TABB
   if(!percenCT){statuT = NAstatus(status)} # no overlap percenCT==0
   # NO STATUS priority on AMBIGUOUS priority on NEW
   elsif(statuT contains "NO STATUS"){status = "NO STATUS"}
   elsif(statuT contains "AMBIGUOUS"){status = "AMBIGUOUS"}
   elsif(statuT contains "NEW"){status = "NEW"}
  } # endif(pA >= ua_Pc)
 } else{print "Format error A"} # endif modeleA
} # endwhile DATA
```

FIG. I.2 – Algorithme pour l'attribution de statuts 'nouveau' et 'ambigu' à certaines CDS uniques aux prédictions *AMIGene*.

Annexe J

Résultats de réannotation

Species Code	Date	Size (Mb)	CDSs				GNF														NG											
			Total		CC		OA -CC		[0-0,2[								[0,2-0,4[		[0,4-0,7[		AP-CC		[0,4-0,7[	[0,7-1]	[0,4-1]							
			OA	AP		/OA		/ OA		/OA	W	W/OA	S	S/OA	NS	NS/OA			/OA			/AP				/AP	N	N/AP	A	A/AP	NS	NS/AP
AERPE	1999	1,67	2694	1721	1545	57,3	1149	42,65	1138	42,24	915	33,96	133	4,94	90	3,34	11	0	0,00		176	10,23	97	42	139	8,08	44	2,56	92	5,35	3	0,17
AQUAE	1997	1,55	1522	1713	1511	99,3	11	0,72	11	0,72	2	0,13	2	0,13	7	0,46	0	0	0,00		202	11,79	124	57	181	10,57	86	5,02	95	5,55	0	0,00
ARCFU	1997	2,18	2436	2459	2360	96,9	76	3,12	65	2,67	19	0,78	23	0,94	23	0,94	10	1	0,04		99	4,03	45	7	52	2,11	21	0,85	30	1,22	1	0,04
BORBU	1997	1,44	851	830	797	93,7	54	6,35	51	5,99	26	3,06	13	1,53	12	1,41	3	0	0,00		33	3,98	20	1	21	2,53	4	0,48	15	1,81	2	0,24
CAMJE	2000	1,64	1647	1620	1617	98,2	30	1,82	25	1,52	9	0,55	8	0,49	8	0,49	5	0	0,00		3	0,19	3	0	3	0,19	0	0,00	3	0,19	0	0,00
CHLPN	1998	1,23	1074	1065	1024	95,3	50	4,66	49	4,56	1	0,09	0	0,00	48	4,47	1	0	0,00		41	3,85	23	1	24	2,25	6	0,56	17	1,60	1	0,09
CHLTR	1998	1,04	893	909	870	97,4	23	2,58	20	2,24	3	0,34	5	0,56	12	1,34	3	0	0,00		39	4,29	18	1	19	2,09	7	0,77	11	1,21	1	0,11
ECOLI	1997	4,63	4289	4100	3959	92,3	330	7,69	321	7,48	61	1,42	91	2,12	169	3,94	9	0	0,00		141	3,44	64	10	74	1,80	30	0,73	42	1,02	2	0,05
HAEIN	1995	1,83	1737	1765	1721	99,1	16	0,92	13	0,75	7	0,40	3	0,17	3	0,17	3	0	0,00		44	2,49	19	4	23	1,30	9	0,51	14	0,79	0	0,00
HELPJ	1999	1,64	1482	1493	1447	97,6	35	2,36	31	2,09	2	0,13	3	0,20	26	1,75	4	0	0,00		46	3,08	30	1	31	2,08	6	0,40	23	1,54	2	0,13
HELPY	1997	1,66	1588	1567	1514	95,3	74	4,66	68	4,28	30	1,89	7	0,44	31	1,95	6	0	0,00		53	3,38	27	3	30	1,91	11	0,70	17	1,08	2	0,13
METJA	1996	1,66	1723	1766	1705	99,0	18	1,04	17	0,99	2	0,12	12	0,70	3	0,17	1	0	0,00		61	3,45	37	5	42	2,38	15	0,85	27	1,53	0	0,00
METTH	1997	1,75	1869	1841	1793	95,9	76	4,07	75	4,01	34	1,82	19	1,02	22	1,18	1	0	0,00		48	2,61	25	0	25	1,36	3	0,16	22	1,20	0	0,00
MYCGE	1995	0,58	483	550	474	98,1	9	1,86	9	1,86	2	0,41	0	0,00	7	1,45	0	0	0,00		76	13,82	37	10	47	8,55	33	6,00	12	2,18	2	0,36
MYCPN	1996	0,81	688	805	664	96,5	24	3,49	23	3,34	2	0,29	6	0,87	15	2,18	1	0	0,00		141	17,52	79	16	95	11,80	37	4,60	56	6,96	2	0,25
MYCTU	1998	4,41	3913	4096	3746	95,7	167	4,27	145	3,71	25	0,64	63	1,61	57	1,46	19	3	0,08		350	8,54	144	13	157	3,83	54	1,32	96	2,34	7	0,17
NEIMA	2000	2,18	2063	1908	1802	87,3	261	12,65	240	11,63	75	3,64	34	1,65	131	6,35	21	0	0,00		106	5,56	36	4	40	2,10	12	0,63	25	1,31	3	0,16
NEIMB	2000	2,27	2128	1960	1810	85,1	318	14,94	308	14,47	124	5,83	46	2,16	138	6,48	10	0	0,00		150	7,65	71	15	86	4,39	50	2,55	32	1,63	4	0,20
PYRAB	1999	1,76	1764	1856	1706	96,7	58	3,29	58	3,29	0	0,00	7	0,40	51	2,89	0	0	0,00		150	8,08	79	11	90	4,85	27	1,45	63	3,39	0	0,00
PYRHO	1997	1,74	2059	1813	1643	79,8	416	20,20	409	19,86	302	14,67	27	1,31	80	3,89	7	0	0,00		170	9,38	90	17	107	5,90	50	2,76	56	3,09	1	0,06
RICPR	1998	1,1	834	886	818	98,1	16	1,92	14	1,68	6	0,72	5	0,60	3	0,36	2	0	0,00		68	7,67	43	8	51	5,76	21	2,37	30	3,39	0	0,00
SYNY3	1996	3,57	3163	3111	2965	93,7	198	6,26	190	6,01	1	0,03	38	1,20	151	4,77	8	0	0,00		146	4,69	61	3	64	2,06	16	0,51	47	1,51	1	0,03
THEMA	1999	1,86	1872	1876	1816	97,0	56	2,99	52	2,78	21	1,12	16	0,85	15	0,80	4	0	0,00		60	3,20	20	6	26	1,39	9	0,48	16	0,85	1	0,05
TREPA	1998	1,14	1040	1034	964	92,7	76	7,31	72	6,92	37	3,56	25	2,40	10	0,96	4	0	0,00		70	6,77	31	0	31	3,00	7	0,68	21	2,03	3	0,29
UREPA	2000	0,75	612	608	589	96,2	23	3,76	22	3,59	3	0,49	7	1,14	12	1,96	1	0	0,00		19	3,13	12	0	12	1,97	0	0,00	12	1,97	0	0,00
VIBCH	2000	4,03	3882	3857	3568	91,9	314	8,09	214	5,51	90	2,32	78	2,01	46	1,18	100	0	0,00		289	7,49	127	3	130	3,37	8	0,21	119	3,09	3	0,08
VIBCH1	2000	2,96	2769	2764	2570	92,8	199	7,19	134	4,84	56	2,02	52	1,88	26	0,94	65	0	0,00		194	7,02	81	0	81	2,93	2	0,07	78	2,82	1	0,04
VIBCH2	2000	1,07	1113	1093	998	89,7	115	10,33	80	7,19	34	3,05	26	2,34	20	1,80	35	0	0,00		95	8,69	46	3	49	4,48	6	0,55	41	3,75	2	0,18
YERPE	2001	4,65	4009	4603	3910	97,5	99	2,47	92	2,29	55	1,37	15	0,37	22	0,55	7	0	0,00		693	15,06	217	11	228	4,95	32	0,70	195	4,24	1	0,02

TAB. J.1 – Comparison of the microbial genes annotated in GenBank files with the CDS predicted by the AMIGene strategy. Abbreviations : CDS, coding sequences ; OA, original annotation ; AP, AMIGA prediction ; CC, common CDS to both OA and AP ; GNF, gene not found ; NG, new gene ; Pc, coding average probability of a CDS ; N, New ; W, wrong ; S, suspicious ; A, ambiguous.

(Ré)annotation de génomes procaryotes complets
Exploration de groupes de gènes chez les bactéries

Résumé court : La stratégie experte semi-automatique de prédiction de Séquences CoDantes (CDS) d'un chromosome procaryote est fondée sur le modèle statistique des chaînes de Markov. Elle est constituée des stratégies *AMIMat* pour l'apprentissage de l'hétérogénéité de composition des CDS d'un chromosome et *AMIGene* pour la reconnaissance et le filtrage des CDS les plus probables. *AMIMat* permet de construire k matrices de transition à partir de k classes de gènes définies selon l'usage des codons synonymes. La précision d' *AMIGene* dépend de la qualité des matrices et d'autres paramètres validés automatiquement par rapport à des annotations de référence. Autour de ces stratégies, un processus de réannotation de génome complet a été développé, en interaction avec notre base multigénome PkGDB, qui facilite l'homogénéisation des annotations des banques. Ce processus de (ré)annotation est utilisé dans de nombreux projets : *Bacillus*, *Neisseria*, *Acinetobacter*, *Entérobactéries*.

Mots-clés : génome procaryote, chaînes de Markov, analyses multivariées, analyse factorielle des correspondances, centres mobiles, hétérogénéité dans l'usage des codons synonymes des CDS, prédiction de gènes, réannotation, exploration d'îlots génomiques.

Prokaryotic complete genome (re)annotation
Exploration of gene groups in bacteria

Short abstract : The semi-automatic expert strategy for predicting CoDing Sequence (CDS) on a prokaryotic chromosome is based on the Markov chain statistical model. It is made up of the strategies *AMIMat* for the training of the compositional heterogeneity of CDS from a chromosome and *AMIGene* for the recognition and the filtering of the most probable CDS. *AMIMat* allows construction of k transition matrices from k gene classes defined according to synonymous codon usage. The *AMIGene* precision depends on the quality of the matrices and on other parameters automatically validated on reference annotations. A complete genome reannotation process, using these strategies, has been developed, in interaction with our multigenome database PkGDB, which facilitates the homogenisation of the databank annotations. This (re)annotation process is used in many projects : *Bacillus*, *Neisseria*, *Acinetobacter*, Enterobacteria.

Key-words : prokaryotic genome, Markov chains, multivariate analyses, correspondence analysis, K-means, CDS synonymous codon usage heterogeneity, gene prediction, reannotation, genomic island exploration.

Atelier de Génomique Comparative UMR 8030–Genoscope
2 rue Gaston Crémieux 91057 CP5706 Evry Cedex France