Phylogénie et évolution des génomes procaryotes

Vincent Daubin

N° d'ordre : 161-2002

Université Lyon 1- Claude Bernard

Thèse

Pour obtenir le grade de

DOCTEUR EN SCIENCES de l'université Lyon 1

Présentée et soutenue publiquement par

Vincent DAUBIN

Phylogénie et évolution des génomes procaryotes

Soutenue le 21 Octobre 2002 devant le jury composé de :

Erick Denamur	Rapporteur
Jean-Pierre Flandrois	Examinateur
Patrick Forterre	Rapporteur
Manolo Gouy	Examinateur
Guy Perrière	Examinateur
Eduardo Rocha	Examinateur

Remerciements (comprenne qui peut...)

Avertissement aux âmes sensibles : *pour avoir lu de nombreux remerciements de thèses en connaissant le derrière des choses, je sais que sous une phrase gentille se dissimule parfois une pointe d'amertume ou d'ironie.*

Ne cherchez rien de tel dans ces pages.

Je suis arrivé au Laboratoire de Biométrie et Biologie Évolutive par un jour pluvieux d'octobre. Tout de suite j'ai pressenti quel calvaire j'allais vivre ici. Les locaux, immenses et déserts résonnaient sous mes pas. Les rares étudiants qui se trouvaient là étaient disséminés aux quatre coins de ce paysage désolé, silencieux, laissant seulement parfois échapper un faible râle d' agonie.

J'ai été littéralement placé sous les ordres de Guy Perrière, un personnage autoritaire et taciturne, mystique illuminé ne buvant que de l'eau, et j'ai rapidement compris que nos goûts et nos caractères étaient diamétralement opposés. Réprimant chacune de mes initiatives, il s'est consciencieusement attaché à déprécier mon travail et à ruiner mon moral pendant ces trois longues années. Je lui tiens une rancune particulière de ne m'avoir jamais permis de présenter mes résultats à la communauté scientifique, que ce soit en France ou à l'étranger. Il en est de même pour la publication d'articles : pas la moindre aide, pas le moindre conseil. J'espère sincèrement ne plus jamais avoir affaire à ce personnage.

J'ai également dû faire face à une adversité redoutable en la personne du directeur de notre équipe, Manolo Gouy. Doté d'un ego surdimensionné, Manolo tâche continuellement d' attirer l' attention à lui en parlant fort et prend un malin plaisir à mésestimer le mérite de ses collaborateurs et étudiants, afin de s'en attribuer tous les lauriers. Des rares discussions que nous avons eu ensemble, où il cherchait à peine à comprendre ce dont je lui parlais, je n'ai retiré que jugements définitifs et sentences destructrices, qui loin de me faire progresser dans ma problématique, m'ont souvent fait me fourvoyer dans des voies stériles. Je ne souhaite à aucun thésard d'avoir quelqu'un comme Manolo pour l'encadrer et si je devais avoir un modèle de chercheur à qui ressembler dans l'avenir, ce n'est certainement pas lui que je choisirais !

Je ne peux évoquer ce trop long séjour au laboratoire sans mentionner le souvenir cuisant de mes contacts avec Laurent Duret. D'un naturel obtus et singulièrement dépourvu

d’imagination, Laurent végète dans son bureau en attendant la retraite. Tous ceux qui le connaissent vous diront qu’une discussion avec lui est immanquablement une perte de temps, et les nôtres n’ont pas échappé à la règle. Et encore faut-il, avec lui, se cantonner à des relations strictement professionnelles, sous peine d’y perdre bien plus que du temps...

Trois ans ! Trois ans à supporter la solitude et le silence pesant de ces locaux, déchirés seulement par les hurlements périodiques et sonores du directeur du Laboratoire, Christian Gautier, expression de son courroux contre quelque étudiant. Si encore j’avais pu trouver du réconfort en dehors de la sphère de la recherche... Mais j’ai dû en particulier supporter, tant au travail qu’en privé, la compagnie récurrente d’un fâcheux, nommé Gabriel Marais, à qui je crains de n’être jamais parvenu à faire comprendre toute l’inimitié qu’il m’inspire. Je me suis souvent heurté à sa vision conservatrice des sciences et de l’évolution, lorsque, sans beaucoup d’illusion, je tentais de l’intéresser à mes résultats ou à n’importe quel autre sujet, et la discussion retombait immanquablement, comme un soufflé. Naïf que j’étais de penser que son avis pourrait m’être d’une quelconque utilité ! Je compte bien ne plus jamais avoir à interagir avec lui, tant du point de vue humain que scientifique. De même, avec les autres étudiants de l’équipe, mes relations ont été constamment tendues et froides. J’ai notamment été plus qu’à mon tour, l’innocente victime de l’humour discutable du sinistre Adel Khelifi.

Je dois terminer en évoquant Emmanuelle, dont j’ai du supporter la présence presque 24h/24 pendant trois ans. Ce fut un véritable enfer d’avoir à tout endosser et de devoir en plus supporter ses crises d’hystérie quotidiennes. Par quel jeu amer du destin nous retrouvons nous de nouveau ensemble, dans le même laboratoire pour l’année qui vient !

Bref, le lecteur aura compris que les conditions dans lesquelles le travail présenté ici a été réalisé, sont parmi les pires que l’on puisse imaginer. Dans l’adversité, j’ai été obligé de faire avec trois bouts de ficelle : le laboratoire, et particulièrement l' équipe de Bioinformatique et Génomique Evolutive, est en effet doté d’un matériel informatique médiocre et les outils qui y sont développés sont pour la plupart sans aucun intérêt, et pas seulement du point de vue de ma problématique.

J’oublie de mentionner bon nombre de personnes à qui je dois des moments difficiles. Je pense qu’elles ne m’en tiendront pas rigueur...

P.S.: si vous n'avez pas tous les éléments pour juger, abstenez-vous de tirer quelque conclusion…

Chapitre introductif : Une brève histoire de la perception du monde procaryote

1 . Chapitre introductif : Une brève histoire de la perception du monde procaryote

1.1 Pléomorphistes contre monomorphistes

Fig. 1.1 : Ehrenberg reconnaît une grande variabilité de formes bactériennes : *Bacterium, Vibrio, Spirochaeta* et *Spirillum*.

L'existence des organismes unicellulaires est connue depuis le XVII^{ème} siècle et les observations de Leeuwenhoek, mais les premières descriptions scientifiques et tentatives de classification de ces « animalcules des infusions » sont le travail d'Ehrenberg dans les années 1830. Il reconnaît un grand nombre de genres d'*infusoria* parmi lesquelles *Bacterium* (bâtonnets droits et rigides), *Vibrio* (bâtonnets tordus non rigides), *Spirochaeta* (filaments spiraux non rigides) et *Spirillum* (filaments spiraux rigides), sans savoir si ce qu'il classifie ainsi représente différentes espèces où différents stades de vie d'un même organisme (Fig. 1.1). Rapidement, ces microbes sont considérés par certains scientifiques comme les représentants d'une seule et même espèce, douée d'une capacité à prendre une grande variété de formes (pléomorphisme), selon leurs conditions de culture ou leur stade de vie. On associe parfois ces organismes au règne animal, mais plus fréquemment à des plantes ou à des champignons dégénérés. Cependant, vers la fin du siècle, un débat s'engage entre les partisans du pléomorphisme, et ceux qui soutiennent l'idée que la diversité des formes observées représente autant d'espèces (monomorphisme). Le débat est contemporain de celui concernant la génération spontanée, et intervient dans un contexte où les idées transformistes commencent à être acceptées par nombre de scientifiques : à la fois la génération spontanée et le pléomorphisme des microbes peuvent être interprétés comme une confirmation des thèses de Lamarck. Les thèses pléomorphistes sont défendues ardemment, notamment par Béchamp qui considère que tout être vivant est constitué de « microzymes » qui peuvent s'assembler en bactéries, qui elles-mêmes peuvent changer de forme. Mais l'histoire a surtout retenu les noms des défenseurs de l'hypothèse

monomorphiste, qui sont en général également « anti-spontanistes », comme Ferdinand Cohn et Louis Pasteur. Parmi eux, Robert Koch, biologiste allemand très attaché à la conception linnéenne de l'espèce, joue un rôle déterminant dans le règlement de la controverse. Il développe la technique de culture bactérienne sur milieu solide (agar-agar) qui permet l'isolation des souches de bactéries. En 1876, il met en évidence que l'agent causal de la maladie de l'anthrax est la bactérie *Bacillus anthracis*, confirmant ainsi la théorie des germes de Pasteur, et en conclura plus tard que chaque forme de cellule observée, chaque maladie correspond à une espèce bactérienne. Malgré une résistance des pléomorphistes, dont certains défendront leur thèse jusque dans les années 1950 (par exemple le Dr. Royal R. Rife, inventeur d'un microscope révolutionnaire dans les années 1930 - voir aussi Wainwright, 1997 pour quelques exemples), la microbiologie entre alors dans une phase de monomorphisme dogmatique, qui permettra d'envisager d'établir une vraie classification. Notamment, en 1844, Hans Christian Gram y contribue de manière importante en décrivant une méthode de coloration qui permet de définir deux groupes de bactéries, dont on découvrira plus tard qu'ils constituent des lignées de grande importance évolutive. Les bactéries sont alors, et pour longtemps, même après l'invention du concept de gène, considérées comme des organismes à reproduction végétative stricte, si primitifs qu'ils ne peuvent posséder des gènes différenciés. Imaginer une sexualité chez ces organismes « pré-géniques » est impossible. Toute variation de formes observée dans une culture bactérienne est attribuée à des contaminations et jette le doute sur le sérieux et la rigueur du manipulateur. Lorsque, au début du XX^{ème} siècle, plusieurs microbiologistes décrivent leurs observations d'une conjugaison bactérienne, ils sont raillés par les monomorphistes (Wainwright, 1997).

1.2 Transfert horizontal et support de l'hérédité : les débuts de la biologie moléculaire

Le premier transfert horizontal décrit est une avancée majeure de la biologie moléculaire, et peut même être vu comme son expérience fondatrice. En 1928, Griffith publie le résultat de ses expériences sur les pneumocoques (Griffith, 1928). Il dispose de deux souches dont l'une est virulente, possède une capsule, et forme des colonies lisses (type S pour Smooth) et l'autre atténuée, sans capsule, forme des colonies rugueuses (type R pour Rough). Les bactéries virulentes, détruites par la chaleur, ne provoquent aucun symptôme chez la souris. Cependant, il observe que lorsque ces débris de bactéries sont injectés en

même temps que la bactérie non virulente, l'animal développe les symptômes de la pneumonie et meurt. Les bactéries récupérées du sang des animaux morts forment des colonies lisses. Griffith en déduit l'existence d'un principe thermostable qui a la capacité de transformer les souches avirulentes en souches virulentes, et ce de manière stable, modifiant donc leur hérédité. Il nomme ce phénomène « transformation ». La question de la nature de ce principe transformant se pose alors. Ce sont Avery, Macleod et McCarty qui, en 1944, mettent en évidence la nature de l'agent de la transformation en tentant de l'isoler par divers traitements (Avery, *et al.*, 1944). Ils observent que seul le traitement par une « désoxyribodépolymérase » est capable de supprimer le pouvoir transformant de la suspension de bactéries lysées par la chaleur. Cette observation est incompatible avec l'hypothèse dominante à l'époque, selon laquelle le gène est de nature protéique. Ils identifient donc le support de l'hérédité comme étant un acide désoxyribonucléique et prédisent que sa structure est plus complexe que l'enchaînement monotone de bases azotées qu'on se représente à l'époque. Malgré le grand soin et la rigueur des expériences d'Avery, Macleod et McCarty, cette découverte est accueillie avec beaucoup de scepticisme et il faut attendre les travaux de Chargaff, Hershey et enfin Watson et Crick pour qu'elle soit pleinement reconnue.

L'expérience de Griffith, outre son impact évident sur les découvertes à venir, constitue la première preuve du fait qu'il existe bel et bien une forme de « sexualité » (dans le sens échange de matériel héréditaire avec d'autres individus) chez les bactéries. Quelques années après la découverte d'Avery, Lederberg et Tatum (Lederberg et Tatum, 1946) décrivent une expérience de complémentation fonctionnelle chez *Escherichia coli* qui montre que les bactéries peuvent échanger des gènes d'une tout autre manière, *via* un mécanisme qui requière le contact physique entre les cellules : la conjugaison. Selon Lederberg lui-même, cette découverte est « postmaturée » (Zuckerman et Lederberg, 1986), c'est-à-dire que le dogme monomorphiste de Koch et Cohn a empêché pendant des années les microbiologistes d'étudier et même d'imaginer l'éventualité d'échanges de matériel héréditaire chez les bactéries.

Les découvertes des années 1940 avaient préparé le terrain pour l'acceptation d'une génétique des bactéries. Outre l'expérience d'Avery, de nombreux résultats de biochimie tendaient à mettre en avant les caractéristiques communes des microbes et des « organismes supérieurs », notamment les expériences de Beadle et Tatum (Beadle et Tatum, 1941) sur le

champignon unicellulaire *Neurospora crassa*, la mise en évidence des propriétés mendéliennes de la transmission des fonctions enzymatiques et la théorie qui en naquit : « un gène = une enzyme ». Ces expériences, dont sont grandement inspirées celles de Lederberg et Tatum chez *E. coli*, établissaient un organisme unicellulaire comme modèle d'étude de la génétique moléculaire.

1952 est une année extrêmement riche pour le sujet qui nous interesse ici : elle voit la mise en évidence par Hayes (Hayes, 1952) du fait que la conjugaison bactérienne consiste en un transfert unidirectionnel d'ADN d'une cellule à une autre et non, comme on pouvait le penser par analogie avec les champignons, en une fusion de cellules. Zinder et Lederberg (Zinder et Lederberg, 1952) montrent que les virus de bactéries (ou bactériophages) sont capables de transporter du matériel génétique de leur hôte et ainsi de participer aux échanges sexuels des bactéries (transduction). Lederberg (Lederberg, 1952) invente d'autre part le terme « plasmide » pour désigner des éléments génétiques extrachromosomiques qui se répliquent de manière autonome. Hershey et Chase (Hershey et Chase, 1952) montrent que seul l'ADN du bactériophage est injecté dans la cellule et qu'il suffit à la multiplication de particules virales dans l'hôte. Enfin, Luria et Human (Luria et Human, 1952) décrivent ce qui sera compris plus tard par Arber et Kehnlein (Arber et Kehnlein, 1967) comme les systèmes de méthylation/restriction de l'ADN dans les bactéries.

Fig. 1.2: « tout ce qui est vrai pour le colibacille est vrai pour l'éléphant »
Dessin de B. Senez 1972

Mon propos n'est pas de décrire ici toutes les avancées de la biologie moléculaire à cette époque. Il convient seulement de noter à quel point l'existence d'une sexualité chez les bactéries a été compliquée à mettre en évidence dans une microbiologie où toute variation était *a priori* soupçonnée d'être une contamination, et l'impact incommensurable qu'a eu cette découverte, *via* notamment les travaux de Jacob, Lwoff et Monod, sur notre vision du monde procaryote avec pour étape ultime l'aphorisme bien connu de Jacques Monod : « Tout ce qui est vrai pour *Escherichia coli* est vrai pour l'éléphant » (Fig. 1.2).

1.3 La systématique moléculaire et la découverte des archées

Au milieu des années 1960, une nouvelle ère s'ouvre pour les évolutionnistes : Zuckerkandl et Pauling (Zuckerkandl et Pauling, 1965) remarquent que les séquences d'ADN et de protéines sont particulièrement bien conservées au cours des temps évolutifs, et qu'elles constituent de ce fait d'excellents marqueurs pour la détection, l'identification et la classification des micro-organismes. Ce n'est que dans les années 1970, avec les travaux de Fox et Woese (Fox, *et al.*, 1977) sur l'ARN de la petite sous-unité du ribosome (16S et 18S), que cette stratégie sera mise en œuvre de manière systématique afin d'établir une classification du monde procaryote. Ces auteurs découvrent alors que la diversité des « procaryotes » a été largement sous-estimé, et suggèrent qu'une division du monde vivant en trois « Urkingdoms » (« royaumes primaires ») est plus appropriée que l'habituelle dichotomie procaryote/eucaryote (Woese et Fox, 1977). Woese fait remarquer que cette dichotomie, définie originellement par Chatton en 1930 et largement considérée comme ayant une base phylogénétique, n'est en réalité qu'une définition par défaut des procaryotes comme non-eucaryotes (Woese, 1987). La dénomination proposée par Woese pour ces trois Urkingdoms (ou domaines) est la suivante : d'une part les Urkaryotes dont la définition équivaut à celle des eucaryotes ; d'autre part, les eubactéries qui représentent la quasi-totalité des bactéries reconnues jusqu'alors ; enfin, les archaebactéries qui, en l'état des connaissances de 1977, semblent toutes présenter la particularité d'être méthanogènes ce qui, étant donné la manière dont on se représente l'atmosphère de la terre primitive, est interprété par Woese comme la preuve de leur ancienneté. L'existence de telles « bactéries » était connue depuis leur description par Sohngen en 1906 (Sohngen, 1906), qui avait montré qu'il existait des bactéries qui pouvaient utiliser le méthane comme source de carbone, et d'autres qui en produisaient. Mais Woese (1977) montre que leur identification comme bactéries n'est due qu'à leur petite taille et que, au niveau de leurs ARNs ribosomaux, qui sont parmi les molécules les mieux conservées du vivant, ces organismes n'ont pas plus de points communs avec les bactéries que ces dernières n'en ont avec les eucaryotes.

Plus tard, Woese propose une phylogénie universelle et une classification du vivant (fig. 1.3) (Woese, *et al.*, 1985; Woese, 1987). Il définit dix divisions majeures parmi les eubactéries sur des critères moléculaires: (1) les bactéries pourpres (« protéobactéries »), (2) les bactéries Gram-positive, (3) les cyanobacteries, (4) les spirochètes et apparentées, (5) les

bactéries vertes sulfureuses, (6) les bacteroïdes, flavobacteries, cytophagales et apparentées, (7) les planctomycetes et apparentées, (8) les Chlamydiales, (9) les micrococcus radiorésistantes et apparentées, et (10) les bactéries vertes non sulfureuses et apparentées (voir tableau 1.1).

Divisions	Subdivision	Genres représentatifs
Protéobactéries	α- protéobactéries	*Agrobacterium, Rickettsia*
	β- protéobactéries	*Thiobacillus, Neisseria*
	γ- protéobactéries	*Escherichia, Legionella*
	δ- protéobactéries	*Myxobacterium*
	ε- protéobactéries	*Helicobacter*
gram-positives	Haut G+C	*Actinomyces, Streptomyces, Mycobacterium*
	Bas G+C	*Bacillus, Clostridium*
	Espèces photosynthétiques	*Heliobacterium*
	Espèces « gram-négatives »	*Megasphaera, Sporomusa*
Cyanobactéries et apparentées		*Nostoc, Synechococcus*
Spirochètes et apparentées	Spirochètes	*Treponema, Borrelia*
	Leptospiras	*Leptonema, Leptospira*
Bactéries vertes sulfureuses		*Chlorobium, Chloroherpeton*
Bactéroïdes, Flavobacteries, Cytophagales et apparentées	Bactéroïdes	*Bacteroides, Fusobacterium*
	Flavobactéries	*Flavobacterium, Cytophaga*
Planctomycetes et apparentées	Groupe des Planctomycètes	*Planctomyces, Pasteuria*
	Thermophiles	*Isocystis pallida*
Chlamydiales		*Chlamydia*
Micrococcus radiorésistants et apparentés	Groupe des Deinococcus	*Deinococcus*
	Groupe des Thermophiles	*Thermus*
Bactéries vertes non sulfureuses	Groupe des Chloroflexus	*Chloroflexus, Herpetosiphon*
	Groupe des Thermomicrobium	*Thermomicrobium roseum*

Tableau 1.1 : Les divisions du domaine des bactéries d'après Woese, 1987 modifié. Woese mentionne l'existence de bactéries non classées dans ces divisions dont notamment *Thermotoga*. Son caractère hyperthermophile et sa position basale dans l'arbre en font un cas particulièrement intéressant.

La plupart de ces groupes ne reposent que très partiellement sur des critères phénotypiques, la variabilité des modes de vie à l'intérieur des divisions étant importante. Par exemple, la division des protéobactéries (« bactéries pourpres »), dont le nom fait allusion à la présence chez certains de ces organismes d'un pigment lié à la photosynthèse, est composée de nombreux groupes (α, β, δ, ε, γ) qui contiennent tous des bactéries dépourvues du fameux pigment, et donc non photosynthétiques. Woese propose que l'ancêtre commun de cette division était photosynthétique et que ce caractère a été perdu plusieurs fois indépendamment (Woese, 1987). Cependant, il note l'extrême hétérogénéité du groupe pour d'autres caractères importants : il existe en effet des protéobactéries hétérotrophes, chimiolithotrophes, anaérobies, aérobies...

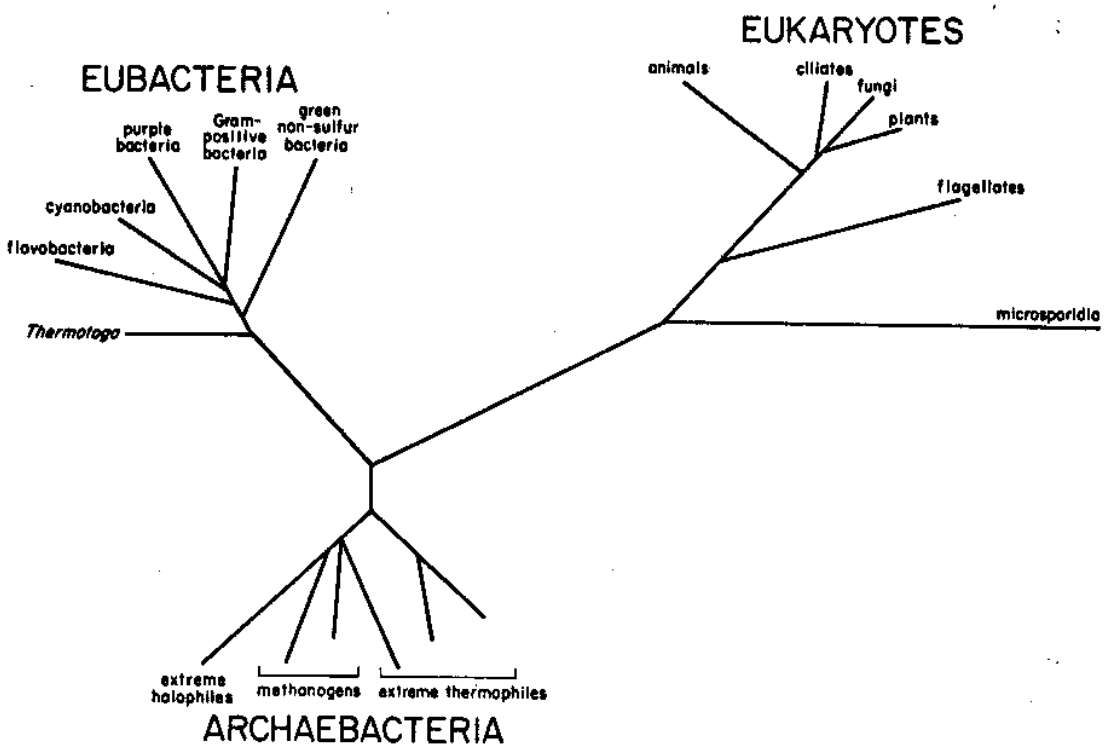


Fig. 1.3 : La phylogénie universelle du vivant basée sur l'ARN ribosomal (Woese, 1987)

Un autre cas est celui de la division des bactéries gram-positives : d'abord définies sur le critère de leur sensibilité à la coloration de Gram, ce groupe s'est avéré contenir un certain nombre de bactéries négatives à la coloration (comme certains mycoplasmes). D'autre part, des bactéries d'autres divisions sont positives à cette coloration comme par exemple *Deinococcus radiodurans*, qui appartient à la division des micrococcus radiorésistantes. En outre, la cohérence de la division des gram-positives a été remise en doute. Il semble en effet qu'elle soit composée de deux grands groupes de bactéries différant par le contenu en bases de leurs génomes et dont la proximité n'est pas certaine : les gram-positives à haut-G+C et les gram-positives à bas G+C (Galtier et Gouy, 1994). Il se pourrait donc que le caractère « Gram-positif » présente un degré assez important de convergence chez les bactéries.

Les relations entre ces grandes divisions sont incertaines dans la phylogénie de l'ARN 16S (Woese, 1987). Les fig. 1.3 et 1.4 montrent que les seuls groupes dont la position est bien soutenue dans l'arbre des bactéries sont les plus basaux. Ces groupes représentent des bactéries hyperthermophiles dont notamment les genres *Thermotoga* et *Aquifex*. L'absence de résolution entre les autres divisions a été interprétée par Woese (Woese, 1987) comme

l'indice d'une radiation, c'est-à-dire une diversification très rapide des phylums bactériens. De nombreuses phylogénies basées sur des protéines présentent également cette absence de résolution entre ces groupes (Koonin, *et al.*, 2001).

A l'époque de la classification de Woese (Woese, 1987), la diversité du monde des archées commence à apparaître : en plus des méthanogènes, on trouve des archées thermophiles extrêmes (hyperthermophiles) ou halophiles. Plus tard, on découvrira que les archées sont présentes dans tous les milieux, et notamment qu'il en existe de très nombreuses qui sont mésophiles. On les subdivisera en deux grands groupes (voir par exemple Brown et Doolittle, 1997) : (1) les Euryarchaeotes contenant des espèces aux caractéristiques écologiques très variables : hyperthermophiles (*Pyrococcus*), méthanogènes (*Methanosarcina*), halophiles (*Halobacterium*), méthanogènes thermophiles (*Methanobacterium*), et (2) les Crenarchaeotes dont la plupart sont hyperthermophiles ou thermoacidophiles (*Sulfolobus, Thermoproteus*). On trouve des mésophiles dans les deux grands groupes. L'existence d'un troisième groupe, les Korarchaeotes, a été proposé sur la base de PCR faites directement sur des échantillons d'eaux de sources chaudes (Barns, *et al.*, 1996) (voir Fig 1.4). Plus récemment, un nouveau groupe d'archées à été découvert dont les membres semblent posséder des tailles de cellule et de génome très réduites : les Nanoarchées (Huber, *et al.*, 2002).Cependant, en 1987, les archées semblent se cantonner à des milieux extrêmes, ce qui conforte l'idée qu'elles conservent à bien des égards des caractères primitifs. Malgré le nom qu'il leur a attribué, Woese n'en fait pas pour autant les représentants de l'ancêtre universel, et préfère voir ce dernier comme un progénote, un organisme « génétique », mais pas encore « génomique », dans lequel ni le nombre de copies d'un gène, ni la spécificité des fonctions qu'il assure ne sont tout à fait fixés (Woese, 1987). De cet ancêtre auraient émergé indépendamment les trois lignées connues aujourd'hui, et les archées évoluant plus lentement et conservant une niche écologique proche de l'ancêtre auraient conservé de nombreuses adaptations aux milieux extrêmes. De même, les bactéries ayant la position la plus basale dans l'arbre, *i.e. Thermotoga* et *Aquifex* auraient hérité leur caractère thermophile de l'ancêtre commun du vivant. Ainsi pour Woese, les différences de longueur de branches observées à la base des trois domaines ne représentent pas le temps écoulé depuis la séparation des lignées, mais le fait que les taux d'évolution ont pu varier entre les lignées, notamment durant la phase de progénote que chacune d'entre elles a dû connaître. Cependant, l'émergence simultanée de ces trois domaines reste hautement spéculative.

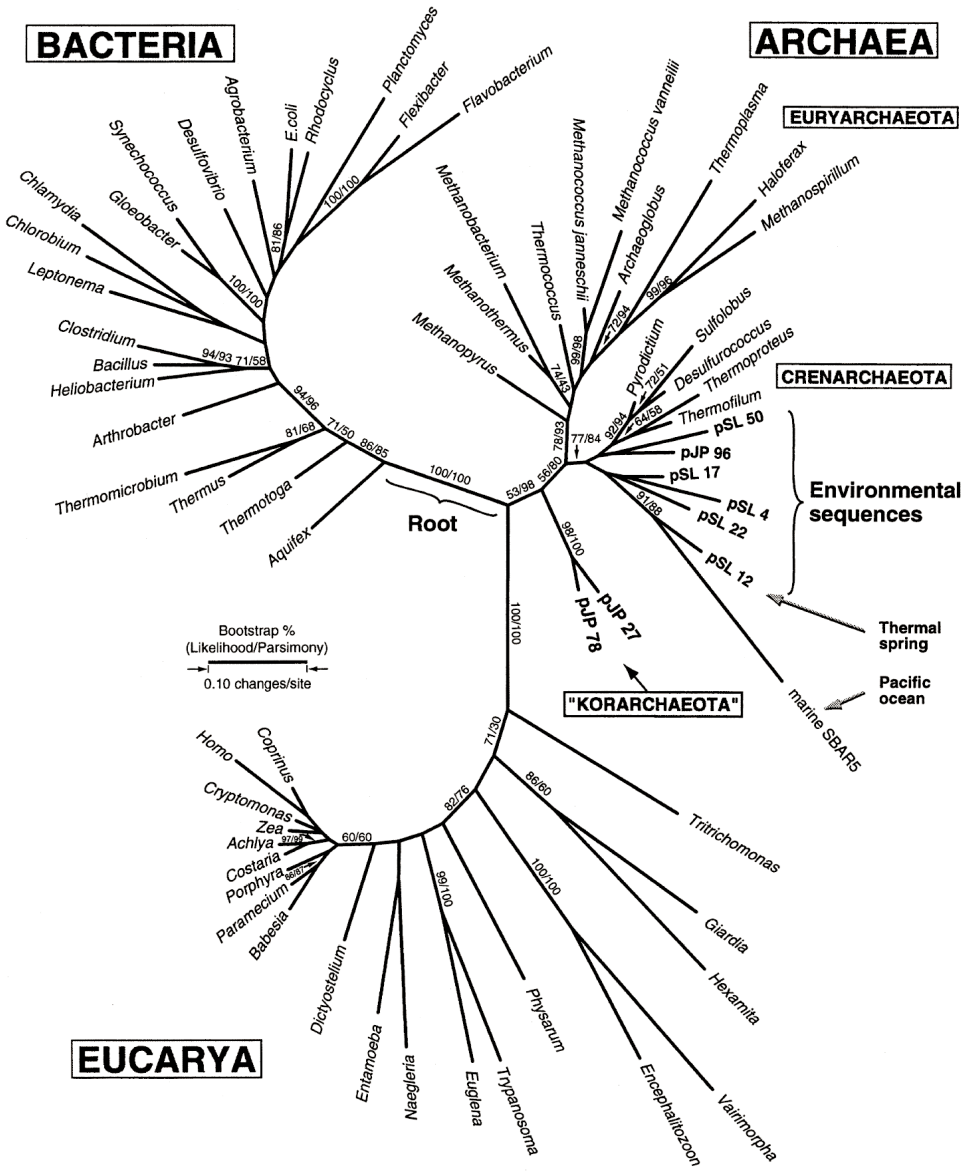


Fig. 1.4 : L'arbre universel du vivant. Basé sur des séquences d'ARN ribosomal et reconstruit avec les méthodes de maximum de vraisemblance et de parcimonie. Seules les valeurs de bootstrap supérieurs à 60% sont indiquées. De nombreuses espèces, et notamment des archées, ont été ajoutée depuis les travaux précurseurs de Fox et Woese (Fox *et al.*, 1977 ; Woese et Fox, 1977). Extrait de Barns, *et al.*, 1996

En 1989, deux articles indépendants proposent de raciner la phylogénie du vivant. Le raisonnement est le suivant : puisque aucun groupe externe, aucune spéciation antérieure à la séparation des trois grands groupes ne peut par définition exister, il faut utiliser un autre type d'événement pour orienter l'arbre phylogénétique. Or, en phylogénie moléculaire, les arbres ne décrivent pas nécessairement la phylogénie des espèces, mais peuvent également permettre de situer, relativement aux événements de spéciation, les événements de duplication du gène considéré. Ainsi, pour raciner l'arbre du vivant, il suffit de trouver une duplication antérieure à la diversification des domaines. Des gènes ayant subi de telles duplications existent : Iwabe *et al.* (Iwabe, *et al.*, 1989) décident d'utiliser les protéines des facteurs d'élongation dont deux formes existent chez tous les organismes : une première dont la fonction est de faciliter la

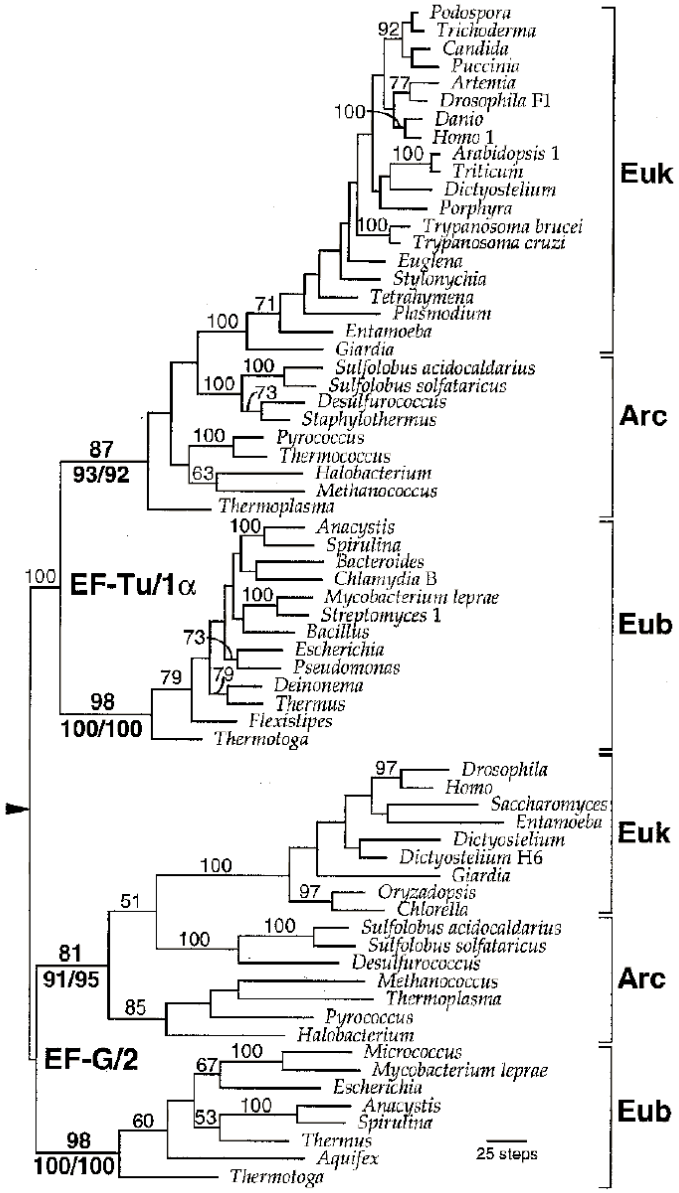


Fig. 1.5 : Phylogénie basée sur les facteurs d' elongation EF-Tu/1α et EF-G/2 racinée réciproquement. L' arbre été obtenu par la méthode de parcimonie. Notez la position de la racine dans la branche des bactéries et la paraphylie des archées dans les deux parties de l' arbre. Iwabe *et al.* (1992) n' ont pas observé la paraphyli des archées visible ici, probablement du fait du faible échantillonnage taxonomique dont ils disposaient. Extrait de Baldauf, *et al.*, 1996.

fixation de l'ARN de transfert chargé au ribosome (EF-Tu chez les bactéries et EF-1α chez les archées et les eucaryotes), et une seconde qui permet la translocation de cet ARN de transfert (EF-G chez les bactéries et EF-2 chez les archées et les eucaryotes) (voir fig 1.5, une phylogénie plus récente des facteurs d'élongation). Gogarten *et al.* (Gogarten, *et al.*, 1989) utilisent quant à eux une duplication précoce qui a donné deux sous-unités de l'ATPase de type V (pour les archées et les eucaryotes) et de type F (pour les bactéries). Ces deux travaux proposent tous deux une racine de l'arbre universel dans la branche des bactéries. Ainsi, les archées seraient le groupe frère des eucaryotes. Cette idée ne provoque pas de grandes surprises car les archées semblent posséder de nombreux mécanismes communs avec les eucaryotes, notamment en ce qui concerne la réplication, la transcription et la traduction. Cette position de la racine est donc rapidement et largement acceptée. Cependant, les données de séquences affluant, la belle image d'un arbre constitué de trois domaines monophylétiques et raciné, par deux phylogénies obtenues indépendamment, dans la branche des bactéries se brouille. D'abord, on découvre l'existence d'ATPases de type V (normalement exclusivement archéennes et eucaryotes) chez des bactéries (Tsutsumi, *et al.*, 1991; Kakinuma, *et al.*, 1991) ainsi que d'ATPases de type F (jusqu'alors uniquement bactérienne) chez une archée (Sumi, *et al.*, 1992), ce qui remet fortement en cause la position de la racine dans le travail de Gogarten *et al.* (Gogarten, *et al.*, 1989). Ensuite, c'est la phylogénie basée sur les facteurs d'élongation qui est remise en cause par Forterre *et al.* (Forterre, *et al.*, 1992), notamment du fait du faible nombre de sites sur lequel est basé l'alignement

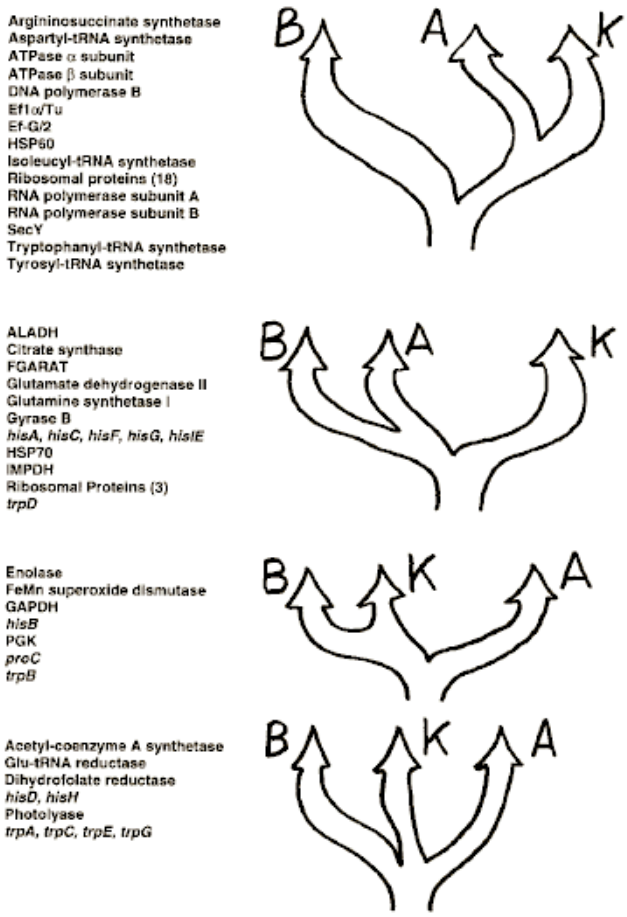


Fig. 1.6 : Les différentes positions de la racine de l' arbre du vivant obtenues en utilisant différente protéines.
B : Bactéries, A : Archées, K : Eucaryotes. Extrait de Brown et Doolittle, 1997

des deux groupes de paralogues. Mais, plus grave, l'accumulation de phylogénies racinées du vivant présentent des résultats contradictoires : à peu près chaque position de la racine imaginable trouve son gène pour la soutenir (voir fig. 1.6). Cependant, il a été montré que pour nombre de ces phylogénies, la position inférée de la racine ne peut être considérée comme correcte du fait de la saturation du signal phylogénétique (Philippe et Forterre, 1999).

D'un autre côté, c'est la monophylie des archées qui est remise en doute. Lake propose dès 1988 qu'un des grands groupes d'archées (les *crénarchées* ou *éocytes*) est plus étroitement apparenté aux eucaryotes qu'aux autres archées. Cette hypothèse est d'abord basée sur une étude de la forme des ribosomes (Lake, 1988) puis sur la découverte d'un insert de 11 acides aminés commun aux eucaryotes et aux crénarchées, et absent chez les euryarchées et les bactéries dans le facteur d'élongation EF-1α (Rivera et Lake, 1992). Elle trouve également un certain soutien dans des phylogénies comme celle des facteurs d'élongation révisée par Baldauf *et al.* (Baldauf, *et al.*, 1996) (Fig. 1.5). D'autres auteurs encore proposent des liens de parenté entre archées et bactéries Gram positives, sur la base du gène HSP70 (Gupta et Golding, 1993; Gupta, 1998a). Pour expliquer ces incongruences entre les différentes phylogénies de gènes, plusieurs hypothèses vont être proposées où le phénomène de transfert horizontal est souvent invoqué, d'une manière ou d'une autre parfois *via* la chimérisation d'organismes.

1.4 Les théories endosymbiotiques

Si l'on fait l'hypothèse que ces phylogénies représentent toutes la véritable histoire des gènes, cela implique que l'évolution des bactéries, archées et eucaryotes à partir de l'ancêtre commun universel est une suite d'événements bien plus complexes que la simple descendance avec modification. Autant de transferts horizontaux entre espèces lointaines peuvent être invoqués pour expliquer ces incongruences, mais l'existence apparente d'un nombre restreint de phylogénies alternatives va suggérer de nouvelles hypothèses.

Dès le XIX^{ème} siècle, les plastes des organismes chlorophylliens ont été soupçonnés d'être des symbiotes. Cependant, c'est seulement dans les années 1960 que cette hypothèse est remise au goût du jour, avec la proposition par Margulis (Margulis, 1970) que non seulement les plastes, mais également les mitochondries constituent les restes

d'endosymbiontes phagocytés par un « protoeucaryote ». Dans leur article de 1977, Fox *et al.* (Fox, *et al.*, 1977) font allusion au fait que l'ARN ribosomal confirme la proximité des plastes et des cyanobactéries. La proximité des mitochondries et des α-protéobactéries apparaît également très clairement dans les premières phylogénies incluant des gènes mitochondriaux (Schwartz et Dayhoff, 1978; Dayhoff et Schwartz, 1981; Schwartz et Dayhoff, 1981). Plusieurs événements indépendants d'endosymbiose ont donc eu lieu de manière certaine au cours de l'évolution des eucaryotes, ce qui semble faire de ce mécanisme un moteur puissant de l'évolution. Chacune de ces endosymbioses a été suivie d'une chimérisation des génomes des protagonistes, provoquant des incongruences phylogénétiques relativement facilement interprétables.

Ainsi, pour interpréter les incongruences observées entre les phylogénies moléculaires, Zillig *et al.* (Zillig, *et al.*, 1985; Zillig, 1987) proposeront que les eucaryotes sont le fruit de la fusion d'une archée et d'une bactérie. Golding et Gupta (Golding et Gupta, 1995), modifiant la thèse de Zillig, proposeront plus tard comme candidat une bactérie gram-négative et une archée éocyte sur la base de l'étude d'un ensemble de 24 phylogénies de gènes dans lesquelles ils décelèrent deux positions concurrentes pour les eucaryotes : l'une correspondant typiquement à celle de l'ARN ribosomal (où chaque domaine est monophylétique et où les distances entre groupes indiquent une proximité des eucaryotes et des archées) et une autre où les eucaryotes étaient significativement groupés avec des bactéries gram-négatives. Selon Golding et Gupta, ce résultat peut s'expliquer par une chimérisation qui se serait située avant l'endosymbiose de l'ancêtre α-protéobactérien de la mitochondrie, et qui aurait provoqué l'apparition du noyau eucaryote. Ces derniers seraient donc le fruit, non pas d'une chimérisation primordiale, mais de deux successives. Cependant, ces résultats ont été critiqués par Roger et Brown (Roger et Brown, 1996) qui attribuent les groupements observés, après réexamen des phylogénies, au choix des séquences utilisées pour reconstruire les arbres. Lorsque toutes les séquences disponibles sont incluses dans l'alignement, plus aucun arbre ne soutient le groupement des eucaryotes et des bactéries gram-négatives. La polyphylie des domaines semble s'expliquer plus rationnellement par de multiples transferts de gènes ou des paralogies non identifiées. D'autres hypothèses plus ou moins semblables à celle de Gupta et Golding ont également été proposées (Cavalier-Smith, 1987, Lake et Rivera, 1994), cependant toutes supposent un événement de fusion ou de phagocytose d'un des partenaires par l'autre, or ni les bactéries ni les archées actuellement connues ne sont capables de phagocyter une cellule si petite soit elle.

Une hypothèse intéressante est, de nouveau, celle proposée par Lynn Margulis (Margulis, 1996), puis reprise indépendamment par Moreira et Lòpez-Garcìa (Moreira et Lopez-Garcia, 1998) et Martin et Müller (Martin et Muller, 1998). S'appuyant sur les associations impliquant des archées et des bactéries observées dans la nature, elle propose non pas une fusion, mais une symbiose intime entre une bactérie, fermentant la matière organique et produisant du H_2 et une archée méthanogène, consommatrice de H_2. Pour Martin et Müller (Martin et Muller, 1998) en particulier, la bactérie impliquée est un α-proteobactérie qui donnera plus tard la mitochondrie. Cependant, l'interprétation des différentes phylogénies reste complexe et implique malgré tout, si l'on suppose que l'histoire des gènes y est réellement représentée, de nombreux transferts horizontaux impliquant des bactéries et des archées.

1.5 Évidences phylogénétiques de transferts horizontaux

> *« From a prokaryotic perspective, sexual eukaryotes like ourselves are incestuous nymphomaniacs: we do « it » too far often and almost exclusively with partners that, from a phylogenetic perspective, are essentially identical to ourselves »*
>
> Levin et Bergstrom, 2000

1.5.1 ***Transferts entre domaines : les relations entre hyperthermophiles***

Des transferts entre bactéries et archées semblent en effet s'être produits au cours de l'histoire des procaryotes. Les plus marquants concernent probablement les bactéries et les archées hyperthermophiles. Déjà suggérés par une étude de Huang et Ito (Huang et Ito, 1999) sur la famille C des ADN polymérases, l'existence de tels transferts déclencha une véritable controverse avec le séquençage complet des génomes d'*Aquifex aeolicus* (Deckert, *et al.*, 1998) et de *Thermotoga maritima* (Nelson, *et al.*, 1999) qui révélaient que respectivement

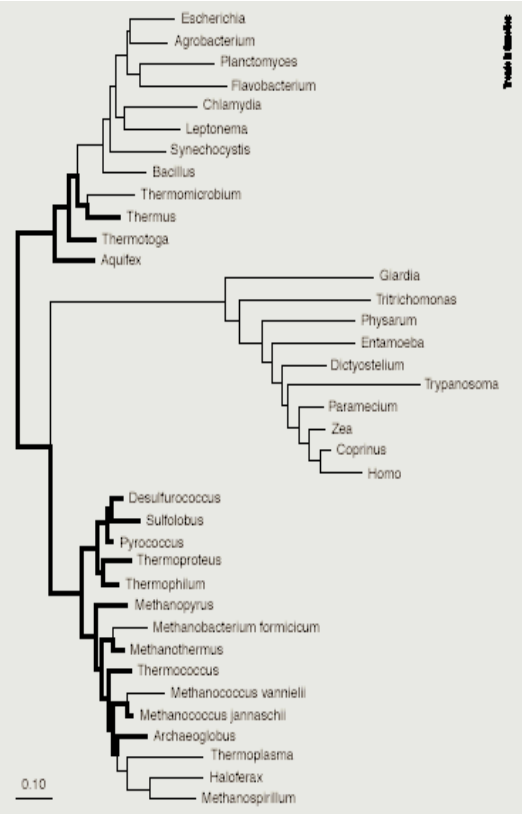


Fig. 1.7 : Les gènes partagés entre hyperthermophiles bactériens et archéens pourraient être l' héritage d' un dernier ancê commun hyperthermophile. La mésophilie serait ainsi un caractère dérivé. La lignée hyperthermophile est symbolisée en gras. Extrait de Kyrpides et Olsen, 1999.

10 % (Aravind, *et al.*, 1998) et 24 % (Nelson, *et al.*, 1999) des ORFs prédites dans ces génomes étaient plus semblables à des gènes archéens qu'à des gènes bactériens. Les méthodes utilisées pour inférer une telle abondance de gènes transférés sont effectivement critiquables : elles se basent sur le meilleur score de BLAST (Altschul, *et al.*, 1997) obtenu pour les ORFs prédites sur les banques de gènes de l'époque. Aravind *et al.* (1998), par exemple, considèrent dans leur analyse du génome d'*Aquifex* qu'un gène a une forte probabilité d'avoir été hérité d'une archée hyperthermophile lorsque la *E-value* (le nombre attendu de « match » au moins aussi bons dans un jeu de données aléatoire) est 100 fois inférieure à celle obtenue chez des bactéries ou des eucaryotes. Kyrpides et Olsen (Kyrpides et Olsen, 1999) font remarquer non seulement que ce critère n'est pas particulièrement stringent, mais qu'en plus la relation entre distance phylogénétique et *E-value* est complexe, et que seule une véritable étude phylogénétique pourrait montrer les véritables liens de parenté entre ces séquences. Ces phylogénies une fois reconstruites ne présentent que rarement un support statistique suffisant pour permettre de conclure au transfert. D'autre part, comme nous l'avons dit plus tôt, dans la phylogénie basée sur l'ARN de la petite sous-unité du ribosome, les bactéries hyperthermophiles *Aquifex* et *Thermotoga* se branchent à la base de l'arbre des bactéries. Cette même position particulière est observée pour les archées hyperthermophiles dans l'arbre des archées. Kyrpides et Olsen (1999) font remarquer que cette position particulière suggère une hypothèse alternative pour expliquer les ressemblances entre les organismes hyperthermophiles : l'héritage de caractéristiques de l'ancêtre commun universel (« Last Universal Common Ancestor » ou LUCA), suivi de pertes ou de fortes divergences des gènes liés à la thermophilie chez les bactéries et archées mésophiles (voir Fig. 1.7). L'idée est séduisante, mais le caractère hyperthermophile du dernier ancêtre commun universel est loin de faire l'objet d'un consensus (Achenbach-Richter, *et al.*, 1987; Forterre, *et al.*, 1992; Miller et Lazcano, 1995; Galtier, *et al.*, 1999;

Glansdorff, 2000; Brochier et Philippe, 2002). En outre, la position phylogénétique de *Thermotoga* et *Aquifex* n'est pas si claire : dans les publications concernant les génomes complets de ces deux bactéries (Deckert, *et al.*, 1998; Nelson, *et al.*, 1999), les auteurs notent qu'en utilisant la grande quantité des gènes désormais à leur disposition, ils n'ont pas réussi à trouver de confirmation significative de la position basale des bactéries hyperthermophiles. Ensuite, Galtier et Lobry (Galtier et Lobry, 1997) ont montré que des contraintes liées à la vie à haute température tendent à enrichir les ARN structuraux en nucléotides C et G chez tous les organismes thermophiles, qu'ils appartiennent aux domaines des bactéries ou des archées. La position d'*Aquifex* et *Thermotoga* dans la phylogénie de l'ARN ribosomal pourrait donc s'expliquer par un biais de composition de ce gène chez les hyperthermophiles. Plus récemment, Brochier *et al.* (2002) ont ré-analysé la phylogénie de l'ARN ribosomal et montré que la position basale des hyperthermophiles était probablement due à la présence de sites évoluant rapidement (fig. 1.8). Dans cette analyse, le groupe des planctomycetes se trouve à la base des bactéries, ce qui pourrait relancer le débat sur la position de la racine de l'arbre du vivant et les caractéristiques de LUCA car ces bactéries possèdent une structure analogue au noyau des eucaryotes. Ceci signifierait une adaptation secondaire de *Thermotoga* et *Aquifex* à la vie à haute température. C'est en effet ce que semblent démontrer les études sur une enzyme qui semble n'être représentée que chez les organismes hyperthermophiles et qui pourrait bien être la condition *sine qua non* de la vie à haute température (Forterre, 2002) : la reverse gyrase (RG). Cette enzyme semble avoir été transférée plusieurs fois indépendamment des archées à certaines bactéries. Ceci suggère que la thermophilie aurait pu n'être « inventée » qu'une fois, *via* la formation du gène de la RG par fusion d'une hélicase et d'une topoisomérase (Confalonieri, *et al.*, 1993) chez l'ancêtre commun à toutes les archées, puis transmise horizontalement, notamment à *Aquifex* et *Thermotoga* (Forterre, *et al.*, 2000).

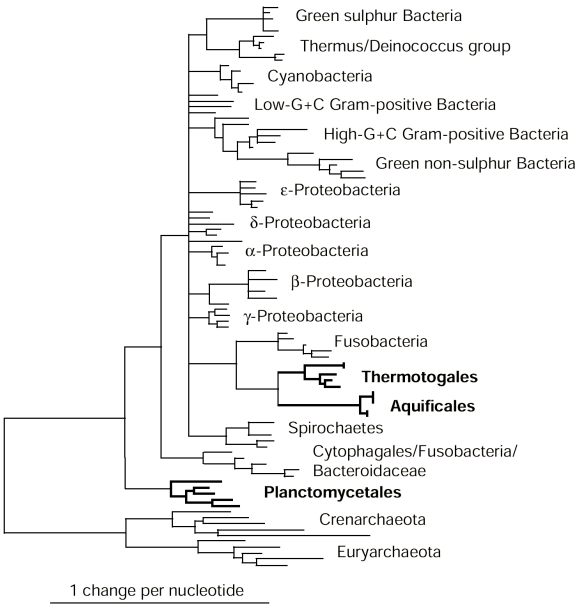


Fig. 1.8 : La position de la racine de l'arbre des bactéries dans la phylogénie de l'ARN ribosomal après élimination des sites évoluant rapidement. Extrait de Brochier et Philippe, 2002.

1.5.2 ***Les transferts horizontaux chez les bactéries***

Dans la phylogénie de l'ARN ribosomal de Barns *et al.* (1996) (Fig. 1.4), si l'on considère que la position basale des hyperthermophiles bactériens peut être artefactuelle, il ne reste plus grand chose de la résolution de la partie bactérienne de l'arbre. Le même problème se pose avec la plupart des gènes utilisés pour inférer une classification phylogénétique des bactéries. Si certains groupes monophylétiques peuvent être retrouvés relativement aisément, comme les protéobactéries ou les cyanobactéries, les liens entre ces groupes restent irrésolus ou contradictoires d'un gène à l'autre dans la plupart des cas. Faut-il supposer que le signal phylogénétique est trop saturé pour permettre de retrouver l'arbre des bactéries ou bien que les transferts de gènes ont brouillé ce signal ?

Jusqu'à récemment, on s'est posé la question de savoir si l'information génétique circulait réellement dans la nature entre souches d'une même espèce bactérienne. Par exemple, Whittam *et al.* (Whittam, *et al.*, 1983) puis Ochman et Selander (Ochman et Selander, 1984), dans une étude du polymorphisme enzymatique des souches sauvages d'*E. coli* à de multiples locus, trouvèrent une remarquable association des différentes formes alléliques et en conclurent que ces populations avaient une structure clonale. Cependant, d'autres observations, notamment entre les souches pathogènes de *Salmonella* montraient une discontinuité de la distribution des facteurs de virulence, suggérant des échanges entre souches (Beltran, *et al.*, 1988).

La vision de populations clonales d'*E. coli* ne fut vraiment contredite que dans les années 1990, notamment par Milkman et Bridges (Milkman et Bridges, 1990; Milkman et Bridges, 1993) qui, utilisant la séquence complète de l'opéron Tryptophane (*trp)* de 36 souches d'*E. coli*, montrèrent que si les relations entre groupes définies par Whittam *et al.* (1983) et Ochman et Selander (1984) ne pouvaient être remises en cause, il existait cependant un certain nombre de régions de l'opéron qui présentaient toutes les caractéristiques d'événements de recombinaison entre les souches des différents groupes. Ces « patrons en mosaïque » montraient clairement l'existence de sous-populations d'*E. coli*, génétiquement distinctes, mais échangeant occasionnellement de l'information par transferts horizontaux.

De nombreux autres cas de transferts horizontaux, basés sur le même type d'observations, ont été décrits au début des années 1990. La plupart impliquent de petites

séquences (< 1kb) contenues dans des gènes ayant un fort impact sur le phénotype de la souche bactérienne receveuse. On peut citer par exemple : le gène de l'endoglucanase *celY* d'*Erwinia chrysantemi*, qui semble être impliqué dans la virulence de ce pathogène de plante (Guiseppi, *et al.*, 1991) ; un gène de capsule d'*Haemophilus influenzae* pathogène (Kroll et Moxon, 1990) ; plusieurs cas de transferts entre souches pathogènes de streptocoques (Simpson, *et al.*, 1992; Whatmore et Kehoe, 1994) ; de nombreux cas de transferts de résistance à des antibiotiques comme la pénicilline chez *Streptococcus pneumoniae* (Dowson, *et al.*, 1993), *Neisseria meningitidis* (Bowler, *et al.*, 1994), *N. gonorrhoeae* (Spratt, *et al.*, 1992) ou la sulfonamide chez *N. meningitidis* (Radstrom, *et al.*, 1992), *etc*…

Groisman *et al.* (Groisman, *et al.*, 1993) montrèrent que le génome des salmonelles a également une structure en mosaïque, en observant la répartition de certaines régions chez d'autres entérobactéries. Ils trouvèrent que plusieurs régions ayant peu ou pas d'homologues chez les autres espèces du groupe avaient une composition en nucléotides C et G très inférieure à la moyenne du génome de *Salmonella*, suggérant qu'elles provenaient de génomes ayant des compositions en bases très différentes. Comme des organismes relativement proches phylogénétiquement (comme les entérobactéries par exemple) ont des taux de G+C comparables, ces gènes devaient nécessairement venir d'organismes plus éloignés. Ces observations s'ajoutent à un certain nombre de cas de gènes ayant visiblement été acquis récemment par *Salmonella* et présentant un faible taux de G+C, comme le gène *phoN* (Groisman, *et al.*, 1992) ou les gènes *rfb* de la synthèse de l'antigène O (Reeves, 1993; Syvanen, *et al.*, 1989). Ces découvertes devaient avoir un impact très important sur l'étude des transferts horizontaux chez les bactéries.

La plupart des transferts décrits précédemment n'impliquent que des bactéries de même espèce. En principe, la fréquence d'intégration d'un ADN dans le chromosome décroît de manière exponentielle avec la divergence de séquence entre les bactéries donneuses et accepteuses (Majewski, *et al.*, 2000). Cependant, l'altération de certaines fonctions cellulaires peut favoriser des échanges entre bactéries plus éloignées par recombinaison homologue. Un cas bien connu et particulièrement important du point de vue évolutif est celui de transferts horizontaux liés à la réparation de gènes dont la défection a été transitoirement sélectionnée, et notamment les gènes du système de réparation des mésappariements (MMR pour « MisMatch Repair »). Des mutants affectés dans les gènes du MMR présentent des taux de mutation particulièrement important et sont appelés « mutateurs ». En outre, certaines de ces

mutations favorisent également des événements de recombinaison avec de l'ADN provenant de bactéries relativement éloignées (comme *Escherichia coli* et *Salmonella typhimurium* par exemple) (Rayssiguier, *et al.*, 1989). Dans des conditions stables, ils sont contre-sélectionnés du fait de l'apparition constante de mutations délétères dans leurs gènes et se maintiennent dans les populations à des fréquences faibles. Cependant, dans des conditions changeantes, leurs taux de mutation et de recombinaison importants peuvent constituer un avantage pour l'exploration de l'espace des allèles possibles, un allèle favorable ayant une plus grande probabilité d'apparaître chez un mutateur. Dans ces conditions, et malgré leur fardeau de mutation, des mutateurs portant un allèle favorable peuvent se fixer dans la population (Tenaillon, *et al.*, 1999). Cependant, si dans cette nouvelle population un non-mutateur apparaît par réversion de la mutation du gène du MMR, celui-ci sera favorisé pour son fardeau de mutation moindre. Le phénotype mutateur n'apporte donc un avantage que transitoirement, pour trouver de nouveaux allèles favorables. Denamur *et al.* (Denamur, *et al.*, 2000) ont montré que la réversion des gènes du MMR semblait se passer très fréquemment par recombinaison avec des allèles d'individus non mutateurs. En effet, de nombreuses incongruences phylogénétiques dans les gènes du MMR peuvent être détectées et témoignent de transferts horizontaux fréquents de petits fragments de gènes (souvent inférieurs à 100 pb) entre souches d'*Escherichia coli.* Il existe une corrélation entre le nombre d'événements de transferts dans un gène et l'importance du phénotype d'hyper-recombinaison dont sa mutation est responsable. Le gène le plus affecté par ces transferts répétés est le gène *mutS* dont l'effet sur la recombinaison est le plus important (Denamur, *et al.*, 2000). Ceci suggère fortement que les événements de recombinaison ont bien eu lieu chez les bactéries ayant le phénotype mutateur correspondant. Ainsi, les mutateurs seraient également des « recombinateurs », c'est-à-dire qu'une mutation comme celles du MMR conduirait à augmenter transitoirement l'adaptabilité *via* les deux processus de mutation et de recombinaison (Tenaillon, *et al.*, 2001).

1.6 Les approches intrinsèques de détection des gènes transférés horizontalement

Sueoka (Sueoka, 1962) a montré qu'il existait une grande diversité des contenus en base G et C des génomes bactériens. Les mycoplasmes peuvent avoir des génomes ne contenant que 25 % de G+C alors que certaines bactéries comme *Micrococcus* peuvent contenir jusqu'à 75 % de G+C. Cette grande variété de contenu en base est due, selon Sueoka, a une « pression de mutation directionnelle » différente d'un organisme à l'autre. Il en résulte que chaque génome a une composition en bases et en oligonucléotides (et notamment en codons) qui lui est propre et qui est considérée comme étant relativement homogène. Un autre facteur affectant la composition des gènes est leur taux d'expression. Gouy et Gautier (Gouy et Gautier, 1982) ont montré que l'usage des codons d'un gène dépendait également de son taux d'expression, et que les gènes d'un organisme pouvaient se regrouper en deux classes selon l'intensité de leur biais d'utilisation des codons : une première classe correspondant aux gènes fortement exprimés (biais fort) et une seconde correspondant aux gènes faiblement exprimés (biais plus faible). Ainsi, pour un gène, le fait d'avoir un usage du code différant à la fois des gènes fortement et faiblement exprimés du génome pourrait être le témoignage d'une adaptation à un précédent génome. Deux approches ont donc été proposées afin d'utiliser cette particularité des gènes acquis récemment pour tenter de les quantifier. La première est due à Médigue *et al.* (Medigue, *et al.*, 1991) qui ont proposé d'utiliser une analyse multivariée de l'usage des codons d'*E. coli*. Utilisant un jeu de séquences représentant près d'un tiers du génome, ils font une Analyse Factorielle des Correspondances (AFC) sur les fréquences relatives des codons et argumentent que les gènes se regroupent non pas en deux classes comme proposé par Gouy et Gautier (Gouy et Gautier, 1982) mais en trois

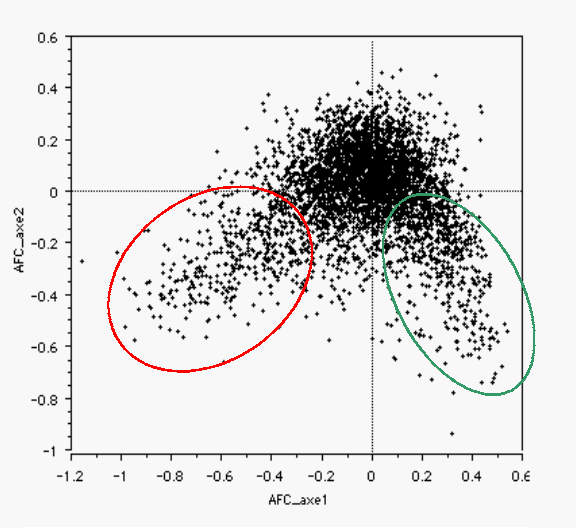


Fig. 1.9: Analyse Factorielle des Correspondances (AFC) réalisée sur les fréquences absolues des codons de 4254 gènes d'*Escherichia coli*. Cette analyse est analogue à celle effectuée par Médigue *et al.* (1991). Les gènes pointés par l' ellipse de droit appartiennent à la classe I (fortement exprimés) et les gènes de l' ellipse d gauche constituent la classe III (gènes transférés horizontalement).

(voir Fig. 1.9) : une première correspondant aux gènes moyennement exprimés, qui représentent la majorité des gènes ; une deuxième contenant des gènes fortement exprimés comme les protéines ribosomales ou les ARNt synthétases ; et une troisième où l'on trouve notamment des plasmides ou des phages. Cette troisième classe est particulièrement intéressante car, pour Médigue *et al.* (1991), elle représente les gènes ayant été acquis récemment par *E. coli.* Cette classe représente plus de 10 % de leur échantillon de gènes, ce qui tend à montrer que les gènes acquis récemment de bactéries très lointaines sont nombreux dans ce génome. Les auteurs notent la richesse en A+T (47 % de G+C en moyenne) des gènes de la 3ème classe en comparaison des deux autres classes (53 % de G+C), ainsi que leur tendance à ne pas éviter les codons rares d'*E. coli* (principalement ATA, AGA et AGG). Une des particularités des gènes détectés comme ayant été acquis par transfert horizontal qui n'est pas discutée par Médigue *et al.* (1991) est leur tendance au regroupement dans l'AFC. En effet, les trois classes sont définies grâce à une méthode statistique de regroupement des points (« clustering ») qui permet de faire une classification en un nombre de classes souhaitées. Si l'on peut facilement argumenter sur des bases biologiques que les deux premières classes constituent des groupes cohérents au niveau de leur usage du code, il est plus hasardeux de le considérer *a priori* pour la troisième. Par définition, des gènes acquis de bactéries phylogénétiquement éloignées devraient former un groupe extrêmement hétérogène. Ainsi, les caractéristiques communes des gènes inférés comme ayant été acquis récemment nécessitent une explication d'ordre biologique. Cet article fut le premier à proposer une détection de gènes transférés horizontalement sans recours à aucune analyse phylogénétique. Beaucoup plus récemment, Moszer *et al.* (Moszer, *et al.*, 1999) proposèrent une analyse du génome de *Bacillus subtilis* avec la même méthode. Les trois mêmes groupes peuvent être identifiés. Dans ce cas également, la 3ème classe (13 % du génome), qui contient des gènes attendus comme fréquemment sujets à des transferts est fortement enrichie en A+T par rapport au génome de *Bacillus* qui possède pourtant un taux de G+C génomique relativement faible (43 % de G+C en troisième position des gènes).

La découverte de Groisman sur les séquences de salmonelles, consolidée par d'autres études (Ochman, *et al.*, 1996; Medigue, *et al.*, 1991) montrant que les gènes acquis récemment possèdent souvent une composition en base différente du G+C moyen du génome (et en l'occurrence souvent plus faible), suggéra que le contenu en G+C, notamment à la position la moins contrainte des codons (la troisième) pouvait permettre de détecter les événements récents de transferts de gènes venant d'espèces lointaines. Ainsi, Lawrence et

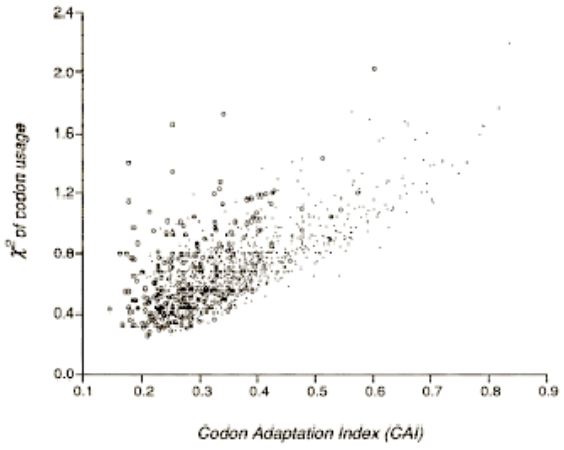


Fig. 1.10: graphe bivarié du CAI et du χ^2 de l' usage du code pour 1189 gènes d'*E. coli*. Les points représentent les gènes natifs (n=1024) et les cercles, les gènes acquis par transfert horizontal (n=165).
Extrait de Lawrence et Ochman (1997).

Ochman (Lawrence et Ochman, 1997) proposèrent d'appliquer cette méthode d'abord à un fragment de séquences représentant près d'un tiers du génome d'*E. coli* (1,43 mégabases soit 1294 gènes) puis au génome complet (Lawrence et Ochman, 1998). Ils utilisèrent trois indices pour détecter les séquences atypiques : le taux de G+C en première et troisième position des codons, le CAI (Codon Adaptation Index - Sharp et Li, 1987) et le χ^2 d'usage du code (sous l'hypothèse d'une utilisation équiprobable des codons) pondéré par la taille des gènes (voir Fig. 1.10). Notamment, ils considèrent que la distribution du taux de G+C en première et troisième position pour les gènes « natifs » doit suivre une loi normale, et que les gènes s'écartant de plus de 2 SE (erreur standard) de la moyenne doivent avoir été acquis récemment (fig 1.11). Ils prédisent ainsi que 17 % du génome d'*E. coli* K12 a été acquis récemment d'organismes éloignés phylogénétiquement et remarquent qu'une proportion de ces gènes plus importante qu'attendue est retrouvée dans la région du terminus de réplication. Comme l'ont noté plus tard Guindon et Perrière (Guindon et Perriere, 2001), ces gènes sont eux aussi beaucoup plus souvent enrichis en A+T par rapport au reste du génome.

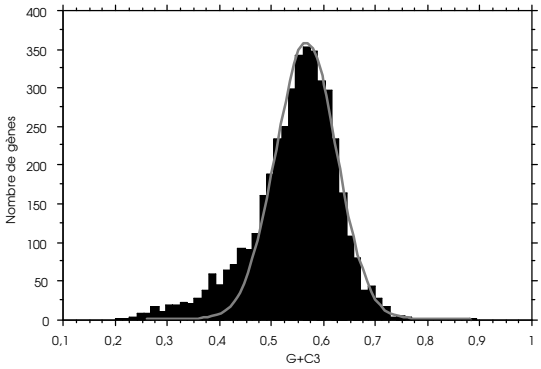


Fig. 1.11 : La distribution du G+C en troisième position des gènes chez *E. coli* et sa comparaison à une loi normale (en gris). Les gènes sortant de cette distribution théorique sont considérés comme ayant été acquis récemment. D'après Lawrence et Ochman, 1997; Lawrence et Ochman, 1998.

Selon leurs auteurs, ces méthodes sous-estiment le nombre de transferts : elles ne sont capables de déterminer des transferts que lorsqu'ils proviennent d'espèce ayant un usage du code différant drastiquement de la bactérie étudiée. Comme il est probable que les transferts horizontaux marchent d'autant mieux entre des espèces relativement proches, le pourcentage de gènes acquis récemment par *Escherichia coli* devrait largement excéder les 20 %. Bien que

certains auteurs comme Syvanen (Syvanen, 1994) remarquèrent très tôt que les approches utilisant la composition des gènes, basées sur des hypothèses fortes, devaient être utilisées avec beaucoup de précaution, le chiffre de 17 % de gènes acquis récemment par *E. coli* est très couramment cité comme un fait avéré.

Species	Number of open reading frames	HGT	Percentage HGT
Proteobacteria			
Escherichia coli	4289	381	9.62
Haemophilus influenzae	1709	96	6.19
Helicobacter pylori 26695	1553	89	6.41
Helicobacter pylori J99	1491	80	5.81
Rickettsia prowazekii	834	28	3.62
Gram-positive bacteria			
Bacillus subtilis	4100	537	14.47
Mycoplasma genitalium	480	67	14.47
Mycoplasma pneumoniae	677	39	5.93
Mycobacterium tuberculosis	3918	187	5.01
Spirochaete			
Borrelia burgdorferi	850	12	1.56
Treponema pallidum	1031	77	8.32
Chlamydiae			
Chlamydia trachomatis	894	36	4.32
Chlamydia pneumoniae	1052	55	5.70
Aquifex aeolicus	1522	72	4.84
Deinococcus radiodurans	2580	95	3.92
Synechocystis PCC6803	3169	219	7.50
Thermotoga maritima	1846	198	11.63
Ureaplasma urealyticum	610	32	5.70

Fig. 1.12 : Nombre de gènes totaux et de gènes détectés comme ayant été acquis récemment par des méthodes basées sur l'usage du code pour différents génomes. Extrait de Garcia-Vallve, *et al.*, 2000.

Le séquençage de nombreux génomes complets ces dernières années a permis de généraliser ce type d'approches basées sur des méthodes intrinsèques. Par exemple, Garcia-Vallvé *et al.* (Garcia-Vallve, *et al.*, 2000) ont créé une base de données accessible sur Internet (*http://www.fut.es/~debb/HGT/*) qui permet de récupérer tous les gènes prédits comme ayant été acquis récemment dans tous les génomes procaryotes disponibles. La méthode utilisée combine un certain nombre d'approches statistiques liées à celles décrites précédemment. Les résultats révèlent une grande disparité entre les espèces bactériennes notamment (voir Fig. 1.12). Le pourcentage inféré de gènes transmis horizontalement chez *E. coli* est inférieur aux précédentes estimations, mais il reste relativement élevé chez des espèces comme *Bacillus subtilis*. D'une manière générale, et étant donné que toutes ces valeurs représentent des sous-estimations, le phénomène de transfert horizontal apparaît ainsi comme un facteur majeur de l'évolution des génomes, et même pour certains auteurs comme le mécanisme roi permettant l'adaptation des bactéries, loin devant la mutation.

1.7 Conjugaison, transduction et transformation : la vie sexuelle des bactéries ?

Il convient ici de faire quelques remarques quant à l'amalgame qui est fait de plusieurs types d'événements dans l'appellation de transfert horizontal. On entend généralement par transfert horizontal, toute acquisition d'ADN qui ne s'est pas faite par la stricte voie verticale de parent à descendant (pour plus de détails sur les mécanismes de transfert d'ADN chez les bactéries, se reporter à l'annexe 1). Ainsi, dans le cas des bactéries, l'acquisition d'un plasmide autoréplicatif constitue un transfert horizontal. De même, l'entrée et la persistance de tout ADN parasite (transposons, bactériophages...) dans la cellule, qu'il s'intègre ou non au génome, est un transfert. Pour ces séquences, la capacité à être transféré est absolument vitale et l'on s'attend à ce qu'elles soient d'une manière ou d'une autre adaptées à ce moyen de reproduction. Enfin, on entend également par transfert horizontal l'intégration dans le génome de gènes dont il n'est pas soupçonné *a priori* que la spécialité est d'être transféré, comme dans la plupart des cas détaillés jusqu'ici.

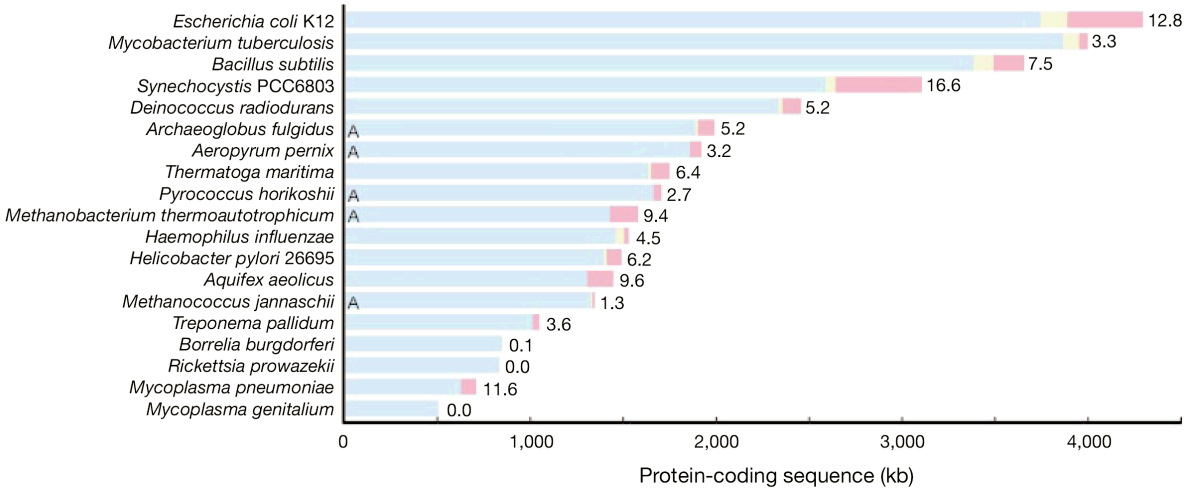


Fig. 1.13 : Représentation graphique de la quantité de gènes « natifs » d'un génome (partie gauche de chaque barre) et des gènes détectés comme ayant été acquis très récemment (partie droite ; pourcentage également indiqué). La partie la plus claire des barres représente la fraction des gènes qui sont associés à des éléments mobiles comme les phages, les plasmides ou les IS (Séquences d'insertion). Extrait de Ochman, *et al.*, 2000

Ainsi, le terme de gène transféré horizontalement peut regrouper des gènes qui sont adaptés au transfert et des gènes qui ne seraient transférés qu'exceptionnellement. Ce sont évidemment ces derniers qui suscitent le plus de débat, particulièrement parce qu'ils

constituent la majorité des gènes détectés comme transférés (voir Fig. 1.13). On imagine la plupart du temps que ces gènes répondent à de fortes pressions de sélection, comme c'est le cas pour les gènes de résistance aux antibiotiques ou aux métaux, de virulence ou encore d'une voie métabolique permettant la survie dans un milieu pauvre. Il a été suggéré dans cette optique que le regroupement des gènes en opéron pouvait être moins un mode de corégulation qu'un moyen pour une voie métabolique complète d'être transmise en une seule fois et ainsi d'augmenter sa probabilité de réussir un transfert (Lawrence et Roth, 1996). C'est en tout cas très probablement cette « pression de transfert » qui est à l'origine des îlots de pathogénicité, regroupement de gènes assurant différentes fonctions liées à la virulence (production d'exotoxines pour détruire les tissus, d'adhésine ou d'invasine pour s'y maintenir et de gènes permettant d'échapper à la réponse immunitaire de l'hôte).

Table 2. Functional Distribution of Genes Proposed as Being Acquired by Horizontal Gene Transfer

Organism	HGT	Info. (%)	Cell (%)	Meta. (%)	Poor (%)	− (%)
Archaeoglobus fulgidus	179	6 (2.2)	22 (8.1)	17 (2.7)	31 (6.0)	103 (14.5)
Aquifex aeolicus	72	7 (3.2)	12 (3.8)	15 (3.6)	20 (6.4)	18 (6.9)
Borrelia burgdorferi	12	2 (1.1)	3 (1.7)	1 (0.8)	3 (2.2)	3 (1.3)
Bacillus subtilis	537	54 (11.5)	34 (5.8)	76 (8.0)	42 (7.3)	331 (21.8)
Chlamydia pneumoniae	55	6 (3.4)	4 (3.0)	10 (4.9)	6 (5.3)	29 (6.8)
Chlamydia trachomatis	36	5 (2.9)	5 (3.8)	9 (4.8)	0 (0.0)	17 (5.7)
Escherichia coli	381	23 (4.8)	41 (6.6)	42 (3.8)	27 (5.4)	248 (15.7)
Haemophilus influenzae	96	3 (1.1)	19 (7.0)	10 (2.2)	3 (1.4)	61 (12.0)
Helicobacter pylori 26695	89	11 (5.0)	3 (1.1)	5 (1.6)	3 (1.6)	67 (11.6)
Helicobacter pylori J99	80	7 (3.2)	6 (2.3)	5 (1.6)	4 (2.2)	58 (11.5)
Mycoplasma genitalium	67	26 (19.1)	10 (16.7)	11 (12.6)	6 (9.1)	14 (10.7)
Methanococcus jannaschii	77	8 (3.6)	11 (7.5)	6 (1.6)	7 (1.8)	45 (7.8)
Mycoplasma pneumoniae	39	8 (5.1)	0 (0.0)	3 (2.7)	2 (2.7)	26 (9.6)
Methanobacterium thermoautotrophicum	179	4 (1.7)	19 (8.8)	26 (5.6)	44 (11.1)	86 (15.5)
Mycobacterium tuberculosis	187	6 (1.6)	20 (4.9)	35 (4.0)	17 (3.2)	109 (6.2)
Pyrococcus horikoshii	154	10 (4.2)	11 (6.1)	8 (2.1)	23 (4.9)	102 (12.7)
Rickettsia prowazekii	28	8 (4.4)	3 (1.9)	12 (6.6)	4 (3.6)	1 (0.5)
Synechocystis PCC6803	219	14 (5.3)	31 (5.2)	15 (2.6)	22 (5.0)	137 (10.6)
Thermotoga maritima	198	13 (5.3)	18 (6.5)	55 (10.7)	47 (12.0)	65 (15.6)
Treponema pallidum	77	8 (4.5)	17 (8.7)	2 (1.4)	17 (10.6)	33 (9.2)

The functional classification available in the COG database (Tatusov et al. 2000) was used. The table shows the number and the group percentage of the genes proposed as being acquired by HGT. The functional groups used were: Info, Information storage and processing; Cell, cellular processes; Meta, metabolism; Poor, poorly characterized; −, not present in any cluster of orthologous group.

Fig. 1.14 : Classification fonctionnelle des gènes détectés comme acquis récemment par Garcia-Vallve, *et al.*, 2000 dans différents génomes. Cette classification est tirée de la banque de gènes homologue COG (Clusters of Orthologous Genes) (Tatusov, *et al.*, 2000). Les gènes annotés « Poor » et « - » ont des fonctions inconnues. Ainsi, chez *E. coli*, le nombre de ces gènes dont la fonction est inconnue est de 275 (sur 381). Extrait de Garcia-Vallve, *et al.*, 2000.

La forte pression de sélection permet ainsi d'expliquer la fixation dans une population d'événements ayant une faible probabilité d'occurrence. Cependant, une pression de sélection forte devrait signifier également un phénotype relativement facile à identifier. Or un grand nombre des gènes pour lesquels on soupçonne un transfert horizontal ont des fonctions encore inconnues (fig. 1.14), ce qui suggère soit que nous n'avons qu'une idée infime des pressions

de sélection qui s'exercent dans les populations naturelles de bactéries, soit que ces gènes sont présents dans le génome pour d'autres raisons encore inconnues. Une explication alternative pourrait être qu'une partie de ces gènes serait des éléments égoïste dont le taux d'insertion serait suffisamment important pour leur assurer un maintien dans les populations.

En outre, les incongruences phylogénétiques mentionnées plus haut concernent souvent des gènes dont la distribution dans le vivant est ubiquitaire. Le transfert horizontal pour ces gènes correspond donc à un remplacement orthologue d'un gène assurant une fonction essentielle et les pressions de sélection favorisant un tel remplacement sont mal comprises.

1.7.1 ***Des caractéristiques communes aux séquences spécialisées dans le transfert horizontal ?***

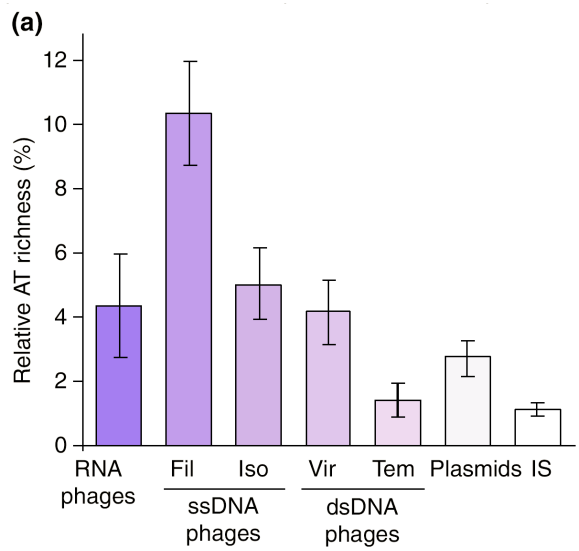


Fig. 1.15 : Richesse relative en A+T des différentes classes de phages, des plasmides et des Séquences d'insertion (IS). Tous ces éléments sont plus riches en A+T que leur génome hôte. Ces résultats sont basés sur l'étude de 52 génomes complets bactériens. Extrait de Rocha et Danchin, 2002

Les bactériophages, IS et plasmides ont généralement des compositions en bases et des usages du code différents des génomes de leurs hôtes. On remarque depuis longtemps que ces séquences ont une tendance à être plus riches en nucléotides A et T que le génome avec lequel elles cohabitent, mais il a été montré récemment que cette direction dans le biais est systématique (Rocha et Danchin, 2002) (fig. 1.15). Ce caractère commun de ces séquences pourrait s'expliquer de plusieurs manières. On peut imaginer que la richesse en nucléotides A et T est un moyen pour l'ADN de favoriser soit son internalisation dans la cellule, soit son intégration dans le chromosome bactérien, ou encore qu'elle permet d'échapper aux mécanismes de restriction présents chez de nombreuses bactéries. Rocha et Danchin (2002) proposent que ce biais de

composition est dû au fait que les molécules d'ATP sont présentes en plus grande concentration dans la cellule, et que l'ADN parasite exploite ainsi mieux les ressources de son hôte. Nous reviendrons plus tard sur ces caractéristiques particulières des séquences sujettes à de fréquents transferts horizontaux.

1.7.2 ***Les bactéries pratiquent-elles le sexe ?***

Le sexe est très fréquemment présenté comme un phénomène général dans le vivant, et l'on entend souvent dire que non seulement les eucaryotes, mais également les bactéries pratiquent le sexe. Chez les eucaryotes, il existe un certain nombre de fonctions qui sont spécifiques au sexe et qui favorisent le brassage des allèles, comme la méiose et la recombinaison qui l'accompagne (Marais, 2002). Par analogie, comme le transfert horizontal semble être un facteur majeur de l'évolution des bactéries, il est souvent supposé (quoique rarement argumenté) que les mécanismes de transfert ont été mis au point pour favoriser l'adaptabilité. Cependant, deux des mécanismes qui permettent aux bactéries d'échanger de l'ADN sont uniquement dus à la présence d'éléments génétiques qui peuvent au mieux être considérés comme des symbiontes, dans le cas où ils transportent des gènes de résistance à des antibiotiques par exemple, mais constituent plus généralement de purs parasites. En effet, les phages et les plasmides conjugatifs transportent avec eux toute la machinerie nécessaire au transfert de gène, et aucune fonction de l'hôte ne semble être spécifiquement impliquée dans ces mécanismes (Levin et Bergstrom, 2000). Les protéines comme IHF (« Integration Host Factor »), dont le nom insiste sur le rôle qu'elle joue dans l'intégration de certains phages dans le génome d'*E. coli,* se révèlent être des protéines participant à la structure du nucléoïde (Dhavan, *et al.*, 2002). On peut donc se poser la question de savoir si les bactéries pratiquent le sexe, au sens où on l'entend chez les eucaryotes, c'est-à-dire si la sélection naturelle a mis en place des mécanismes favorisant le brassage des allèles des différents gènes.

La transformation semble ainsi être le seul mécanisme qui puisse avoir été élaboré pour remplir cette fonction : les bactéries possèdent pour la plupart des mécanismes actifs d'internalisation de l'ADN libre, et nombre de gènes ont été identifiés au départ comme étant spécifiquement impliqués dans les mécanismes de recombinaison (comme les gènes *Rec* ou *Ruv*) (Cox, 2001; Lusetti et Cox, 2002). De plus, certaines séquences, comme les séquences *chi* d'*E. coli* ont été décrites comme étant fortement recombinogènes, et interprétées comme

adaptées à la réparation des gènes ou au remplacement d'allèles entre souches d'une même espèce. Cependant, des travaux récents contredisent assez fortement cette vision. D'abord, on peut remarquer que la compétence est sujette à régulation chez de nombreux organismes comme *Bacillus subtilis, Streptococcus pneumoniae* ou *Haemophilus influenzae*. Si la compétence est avant tout un moyen de réparer l'ADN, on doit observer une induction des gènes de compétence lorsque l'ADN est endommagé comme c'est le cas pour tous les autres mécanismes de réparation de la cellule. Redflied (Redfield, 1993), a montré que ce n'était pas le cas chez *B. subtilis* et *H. influenzae*. Chez plusieurs bactéries compétentes, l'induction de certaines enzymes impliquées dans les mécanismes de réparation, comme RecA a été interprétée comme une adaptation favorisant la recombinaison, mais il a également été proposé que l'entrée d'ADN simple brin dans la cellule serait responsable d'un faux signal d'endommagement de l'ADN (Redfield, 2001). De plus, il parait surprenant dans l'hypothèse d'une fonction dans la réparation de l'ADN que chez de nombreuses bactéries, la compétence soit induite pendant ou a la fin de la phase exponentielle de croissance (Hahn, *et al.*, 1996; Echenique, *et al.*, 2000; Macfadyen, 2000; Berka, *et al.*, 2002). Macfadyen *et al.* (MacFadyen, *et al.*, 2001) ont récemment montré que la compétence était inhibée chez *H. influenzae* par la présence de nucléotides ou de nucléosides puriques dans le milieu. Ce mode de régulation ressemble plus à celui attendu pour les gènes d'une fonction nutritive. Macfadyen *et al.* (2001) proposent que la compétence constitue avant tout un moyen d'obtenir des nucléotides du milieu environnant. Certains faits cependant restent inexpliqués sous cette hypothèse, et notamment le fait que certaines bactéries comme *Neisseria meningitidis* ou *H. influenzae* possède un mécanisme de reconnaissance de l'ADN à interner qui favorise l'entrée de séquence de la même espèce.

D'un autre coté, les gènes connus pour être impliqués dans la recombinaison (comme les protéines des voies *Rec* et *Ruv*) se sont révélé être plus spécifiquement des gènes de la réparation associés à la réplication de l'ADN (« recombinational repair ») (Cox, *et al.*, 2000). La fonction de ces gènes dans la réparation des lésions et la résolution des fourches de réplication leur permettrait d'intervenir également, mais presque de manière anecdotique, dans la recombinaison entre souches. De même, la fonction des sites *chi* d'*E. coli* se révèlerait être plus d'orienter le mécanisme de réparation de l'ADN impliquant *RecBCD* au niveau de la fourche de réplication (Kuzminov, 1995; Horiuchi et Fujimura, 1995) que de favoriser la recombinaison.

Rosemary Redfield (Redfield, 2001), au regard de ces faits, a récemment défendu la thèse selon laquelle les bactéries ne feraient pas de sexe dans le sens où aucun gène n'aurait été sélectionné spécifiquement chez les bactéries pour favoriser les échanges d'ADN. La raison pour laquelle aucun mécanisme n'a été sélectionné serait qu'à l'instar de la mutation, le transfert d'ADN ne serait qu'exceptionnellement bénéfique, et même le plus souvent très dommageable à la cellule. Selon ce point de vue, de même qu'il existe des mutateurs qui peuvent être sélectionnés de manière transitoire dans des conditions de stress, des bactéries ayant la capacité d'intégrer de l'ADN étranger dans leur génome pourraient être temporairement avantagées, mais les inconvénients de ce système seraient trop important pour que des fonctions spécifiques d'incorporation d'ADN dans le génome soient sélectionnées à long terme.

1.8 Du clone à la chimère

> *« If « chimerism » or « lateral gene transfer » cannot be dismissed as trivial in extent or limited to special categories of genes, then no hierarchical universal classification can be taken as natural. Molecular phylogeneticists will have failed to find the « true tree, » not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree. »*
>
> Doolittle, 1999b

Deux grands types de résultats tendent à mettre en évidence des transferts horizontaux. Nous avons vu que d'une part un nombre important d'incongruences phylogénétiques suggèrent des transferts impliquant des organismes très éloignés et que d'autre part, un pourcentage élevé des gènes des génomes complètement séquencés présentent des

compositions en bases atypiques. La connexion entre ces deux types de résultats n'a pu être démontrée que rarement. Un test récent de la capacité des différentes méthodes de détection des transferts horizontaux à retrouver les mêmes gènes a montré notamment que méthodes phylogénétiques et méthodes basées sur la composition ont tendance à identifier des ensembles de gènes distincts (et même de manière surprenante, plus distinct qu'attendu par hasard) (Ragan, 2001). Malgré cela, il est tentant d'englober l'ensemble de ces résultats dans un modèle d'évolution des bactéries où les échanges entre « espèces » éloignées sont la règle. Plusieurs auteurs s'y sont attaché, dont notamment W. F. Doolitlle (Doolittle, 1999b; Doolittle, 1999a) qui propose que l'évolution des procaryotes serait plus fidèlement représentée par un réseau que par la traditionnelle métaphore de l'arbre (fig. 1.16). Selon cette hypothèse, il serait vain de tenter de reconstruire l'histoire des espèces bactériennes et seules les histoires des gènes nous seraient accessibles. Dans le même ordre d'idée, William Martin (Martin, 1999) et d'autres (Bellgard, *et al.*, 1999) suggèrent que l'acceptation du très fort taux de transferts horizontaux pourrait permettre de mieux comprendre les grands principes qui gouvernent la distribution des gènes au sein des génomes, ou de « l'espace génomique » procaryote. Ainsi, non seulement les eucaryotes, mais également, et dans une plus large mesure, les procaryotes seraient des organismes chimères. Il est amusant de noter que la notion d'espace génomique (Bellgard, *et al.*, 1999) rappelle dans une certaine mesure le continuum de formes bactériennes postulé par les pléomorphistes du milieu du XIX^{ème} siècle. Ainsi, de la Cruz et Davies (de la Cruz et Davies, 2000) n'hésitent pas à affirmer:

« It is clear that genes have flowed through the biosphere, as in a global organism. HGT, once solely of interest for practical applications in classical genetics and biotechnology, has now become the substance of evolution. »

Sous cette hypothèse, la force majeure expliquant les relations entre procaryotes serait la structuration du milieu, et les opportunités qu'ils ont d'échanger des gènes. Doolittle envisage ainsi que les relations phylogénétiques que l'on infère entre les espèces pourraient ne représenter que leur propension à échanger régulièrement des gènes (Doolittle, 1999a).

Cependant, d'une part la présence de séquences atypiques dans les génomes et d'autre part les incongruences phylogénétiques observées peuvent trouver d'autres types d'interprétation que le seul transfert horizontal. Dans cette optique, il serait particulièrement

dramatique pour notre vision du monde procaryote, d'attribuer à des transferts des observations dont l'explication pourrait être un simple artefact méthodologique.

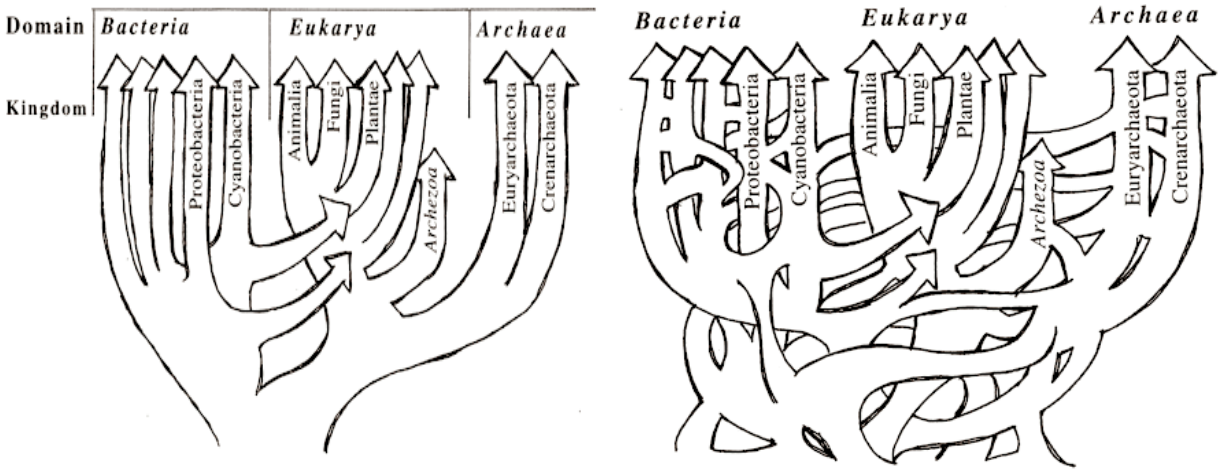


Fig. 1.16 : De l' arbre au réseau : l' évolution de la perception du monde procaryote selon Doolittle (Doolittle, 1999b)

Plusieurs travaux récents critiquent l'utilisation du transfert horizontal comme explication systématique des patrons phylogénétiques et compositionels atypiques (voir notamment Kurland, 2000; Guindon et Perriere, 2001; Wang, 2001). Les arguments sont nombreux : d'abord, en ce qui concerne les phylogénies, les incertitudes sur les branchements anciens sont souvent minimisées et de nombreux cas de transferts supposés pourraient n'être que des sur-interprétations des arbres ; ensuite, certaines hypothèses sur lesquelles sont basées les méthodes utilisant la composition des gènes, notamment la faible hétérogénéité intrinsèque des génomes bactériens, pourraient bien ne pas être vérifiées. Ces points seront abordés dans les chapitres suivants.

Chapitre 2 : Approche phylogénomique et transferts horizontaux chez les procaryotes

2 Chapitre 2 : Approche phylogénomique et transferts horizontaux chez les procaryotes

Malgré le succès grandissant de la vision d'un « organisme global », certains évolutionnistes ne désespèrent pas de reconstruire l'histoire évolutive des procaryotes. Si aucun gène ne peut être considéré *a priori* comme représentant la phylogénie des espèces à lui seul, peut-être la mise en oeuvre de méthodes se basant sur des niveaux supérieurs d'organisation peut-elle pallier ce problème. De nombreuses méthodes ont été proposées à cette fin. Elles peuvent être regroupées en deux classes : les méthodes se basant sur la ressemblance globale des génomes (notamment le contenu en gènes), et celles qui utilisent des alignements de gènes concaténés.

Parallèlement, mais souvent de manière disjointe, des tests systématiques de la congruence des données phylogénétiques ont été tentés pour déterminer si des gènes partagent une même histoire évolutive ou bien si les relations entre espèces sont plus fidèlement représentées par des réseaux d'échanges de gènes.

Je vais détailler dans ce chapitre le principe et les résultats de certains de ces travaux.

2.1 La phylogénie à l'heure de la génomique.

Nous allons examiner comment les évolutionnistes moléculaires ont proposé d'utiliser les masses de données phylogénétiques disponibles afin de retracer l'histoire des bactéries. Pour quelques explications sur le vocabulaire phylogénétique utilisé dans ce chapitre, se reporter à l'annexe 2.

*2.1.1 **Concaténer les gènes***

La méthode la plus intuitive pour tenter de résoudre les problèmes rencontrés avec les phylogénies de gènes, est la méthode de concaténation. En effet, si un alignement ne contient

pas assez d'information sur les relations entre espèces éloignées, la multiplication des données, *via* la construction de super-alignements devrait permettre d'augmenter significativement cette information. La restriction majeure de cette approche est la disponibilité d'un nombre réduit de familles de gènes ubiquitaires : en effet, concaténer des gènes qui sont absents chez certaines des espèces étudiées pose le lourd problème de gérer les données manquantes dans les alignements de séquences.

Idéalement dans une telle approche, même si des gènes ont subi des transferts horizontaux, l'information aberrante qu'ils apportent sur certaines parties de l'arbre devrait être diluée par les informations congruentes qu'apportent les autres gènes. Cependant, cette propriété n'est vraie que si les gènes aberrants sont peu nombreux comparativement aux autres gènes. Un premier travail de Teichmann et Mitchison (Teichmann et Mitchison, 1999) conclut de manière plutôt pessimiste sur ce dernier point. Utilisant 32 gènes protéiques concaténés, ces auteurs essayent de retrouver les relations phylogénétiques de sept bactéries et deux archées. Ils obtiennent un arbre robuste qu'ils jugent artefactuel, notamment du fait de la position basale de la bactérie spirochète *Borrelia burgdorferi*. Après une analyse individuelle plus poussée de chacun des 32 gènes, ils trouvent que les spirochètes possèdent un gène d'origine archéenne, celui de la chaîne β de la phénylalanyl-tRNA synthétase. Deux autres gènes leur semblent suspects de transferts horizontaux ce qui les conduit à réduire leur jeu de séquences à 29 protéines concaténées. L'arbre résultant est nettement moins robuste que le premier et présente certains regroupements qui sont probablement dus à des artefacts d'attraction des longues branches. Cette expérience montre que même un nombre restreint de gènes ayant subi un transfert peut affecter fortement la topologie de l'arbre. D'autre part, Teichmann et Mitchison (1999) concluent que les méthodes de phylogénie moléculaire, y compris la méthode de concaténation, sont probablement incapables de résoudre les relations de parenté entre des groupes aussi anciens que les différents phylum bactériens, du fait d'une trop grande saturation du signal phylogénétique.

Brown *et al.* (Brown, *et al.*, 2001) ont plus récemment effectué une étude très semblable, mais en utilisant 23 gènes partagés chez 45 espèces. Leur premier arbre est très robuste et présente étrangement le même branchement considéré comme aberrant que dans l'étude de Teichmann et Mitchison (1999) : les spirochètes ont la position la plus basale chez les bactéries. Encore une fois, les auteurs interprètent cette topologie particulière par la présence de gènes ayant subi des transferts horizontaux dans l'alignement. Après analyse des

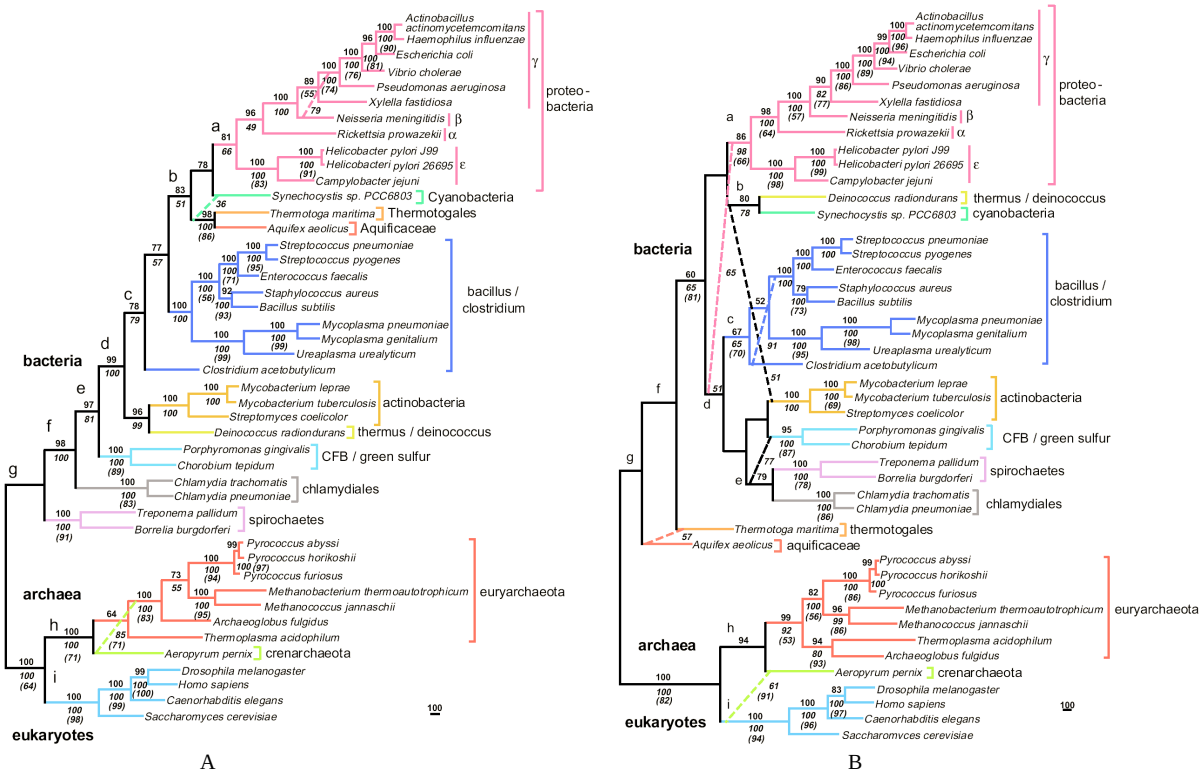


Fig. 2.1: l'arbre du vivant (45 espèces) basé sur un jeu de protéines concaténées. A : l'arbre basé sur 23 protéines concaténées. B : l'arbre après la suppression des neuf gènes suspects de transferts horizontaux. Les pointillés représentent les groupements alternatifs lorsque les différentes méthodes de reconstruction ne donnent pas le même résultat.
Extrait de Brown, *et al.*, 2001.

phylogénies une par une, neuf alignements sont exclus (dont celui de la chaîne β de la phénylalanyl-ARNt synthétase), soupçonnés d'avoir subi des transferts entre domaines. La nouvelle phylogénie obtenue est plus en accord avec celle basée sur l'ARN ribosomal. On peut remarquer que la partie bactérienne de l'arbre, si elle est considérée seule, n'a que très peu changé d'un point de vue topologique entre les deux arbres: la modification réside essentiellement en une rotation de la partie bactérienne vis à vis de la racine, qui passe ainsi de la branche des spirochètes à celle plus consensuelle des bactéries hyperthermophiles. Cependant, elle est nettement moins soutenue et présente des différences importantes selon la méthode de reconstruction utilisée. Ainsi, les résultats de Brown *et al.* (2001) vont à première vue dans le sens des conclusions de Teichmann et Mitchison (1999) quant à la capacité des méthodes classiques de phylogénie moléculaire à résoudre les problèmes de phylogénie profonde, puisqu'une fois les gènes suspects supprimés de l'alignement, les nœuds profonds sont peu résolus.

Ce travail appelle plusieurs remarques : d'une part, il semble montrer que de très nombreux gènes ubiquitaires (9 sur 23 soit près de 40 %), assurant des fonctions essentielles ont subi des transferts horizontaux inter-domaines, notamment entre bactéries et archées. Ces transferts ont un impact fort sur la topologie de l'arbre. Le seul critère sur lequel Brown *et al.* (2001) se sont basés pour inférer des transferts horizontaux est la non monophylie des bactéries dans les arbres basés sur les gènes individuels. Ainsi, si des transferts ont eu lieu entre espèces du même domaine, ils n'ont pu être détectés. Or si 40 % des gènes ont subi des transferts inter-domaines, on peut envisager que de nombreux transferts intra-domaines (théoriquement beaucoup plus probables) perturbent également la topologie de l'arbre.

Une autre remarque qui peut être faite concerne l'importance donnée à chacun des gènes. L'alignement constitué des 14 protéines contient 3824 acides aminés, mais seulement quatre de ces protéines représentent plus de la moitié des sites. Ainsi, si l'une de ces protéines soutient une topologie aberrante du fait d'un transfert, d'une paralogie cachée ou d'un artefact de reconstruction, il est probable que son impact sur la topologie finale sera très important.

Enfin, la méthode utilisée chez Brown *et al.* (2001) ainsi que chez Teichman et Mitchison (1999) pour identifier les gènes ayant subi des transferts est discutable. En effet, l'approche de concaténation se justifie lorsque l'on admet que, pour une raison ou une autre, les arbres individuels peuvent être faux (dans le sens où ils ne représenteraient pas la phylogénies des espèces), tout en contenant une information qui pourra être mise en évidence par celle apportée par les autres gènes. La sélection opérée par Brown *et al.* (2001) est en désaccord avec ce principe. Il est en effet contradictoire de considérer que l'approche de concaténation peut corriger les défauts des phylogénies de gènes individuels dans le groupe des bactéries et pas dans la phylogénie globale. Le fait d'enlever des gènes sur ce critère peut en effet diminuer l'impact des transferts inter-domaines, mais donne d'autant plus de poids à d'éventuels transferts intra-domaine. La mise en place de ce critère est d'autant plus dommageable que le nombre de familles disponibles est réduit. De ce point de vue, il pourrait être intéressant d'introduire dans la méthode une recherche moins biaisée des incohérences entre alignements à concaténer. Nous reviendrons plus tard sur ce point (voir section 2.5).

2.1.2 ***Mesurer la ressemblance globale entre génomes***

2.1.2.1 Le contenu et l'ordre des gènes

Du point de vue de Teichmann et Mitchison (1999), le seul moyen de résoudre des problèmes de phylogénie profonde est d'envisager des méthodes moins sensibles à la saturation du signal phylogénétique que la phylogénie moléculaire. Dans le contexte du nombre grandissant de séquences de génomes complets procaryotes, de nouveaux types de caractères informatifs sont disponibles.

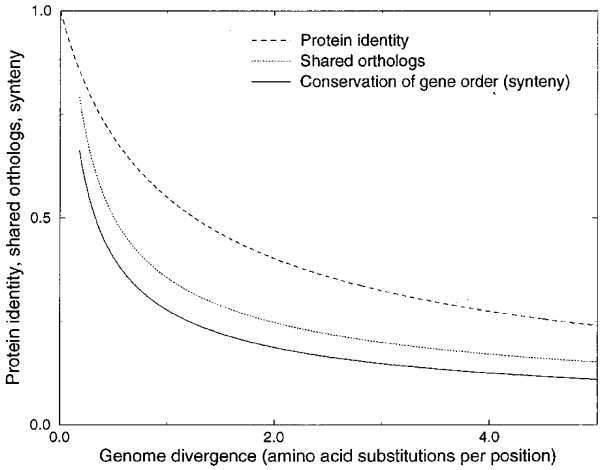


Fig. 2.2 : Taux relatifs d'évolution des génomes. En abscisse, la divergence des protéines estimées sur un jeu de 34 orthologues. Les courbes concernant les orthologues partagés et la conservation de l'ordre des gènes ont été ajustées aux données de neuf génomes complets.
Extrait de Huynen et Bork 1998.

Huynen et Bork (Huynen et Bork, 1998) montrent qu'au moins deux types d'information autres que la divergence des séquences peuvent être utilisés pour mesurer l'évolution des génomes. La première d'entre elles est la fraction de gènes orthologues partagée par deux génomes. La seconde est la conservation de l'organisation des gènes sur le chromosome (synténie). Ces deux mesures reflètent assez bien la divergence entre génomes, bien qu'elles apparaissent moins efficaces que les mesures plus traditionnelles comme l'identité entre protéines, notamment pour des espèces éloignées (Fig. 2.2). Il convient ici de s'attarder sur quelques définitions : à l'origine, le concept d'orthologie est basé sur une définition phylogénétique (sont orthologues deux gènes homologues qui ont acquis leur indépendance évolutive après un évènement de spéciation) et s'oppose à celui de paralogie (sont paralogues deux gènes homologues qui ont acquis leur indépendance évolutive après un évènement de duplication) (Fitch, 1970). Il est important de noter que contrairement à ce que certains auteurs ont cru comprendre de l'article de Fitch (1970) (voir par exemple les instigateurs de la méthode de meilleure similarité

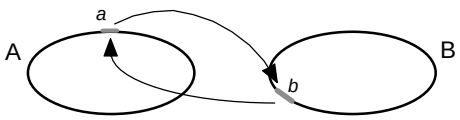


Fig. 2.3 : Définition de l'orthologie par meilleure similarité réciproque. *a* et *b* sont orthologues si, dans le génome B, *b* a le meilleur score de similarité avec *a* et réciproquement.

réciproque Mushegian et Koonin, 1996), aucune hypothèse sur la fonction des gènes n'intervient dans cette définition. Ainsi, seule la reconstruction de l'histoire d'une famille de gènes devrait permettre d'inférer ce type de relation. Cependant cette définition est difficilement applicable à grande échelle, et un moyen plus pratique a été proposé (Mushegian et Koonin, 1996), celui du meilleur score de similarité réciproque (Fig. 2.3). Bien que la définition de l'orthologie sur des critères de similarité puisse conduire à de fausses prédictions dans certains cas particuliers (voir section 2.1.2.3), elle est largement considérée comme donnant une bonne approximation des relations d'homologie entre gènes.

De même, le concept de conservation de l'organisation des gènes (synténie) chez les bactéries est compliqué à mettre en oeuvre du fait notamment que tous les gènes sont généralement sur le même chromosome circulaire et que l'ordre des gènes est très mal conservé lorsqu'on considère des espèces même faiblement éloignées. La mesure proposée par Huynen et Bork (Huynen et Bork, 1998) est en fait basée sur le nombre de paires de gènes (gènes adjacents) conservées entre espèces.

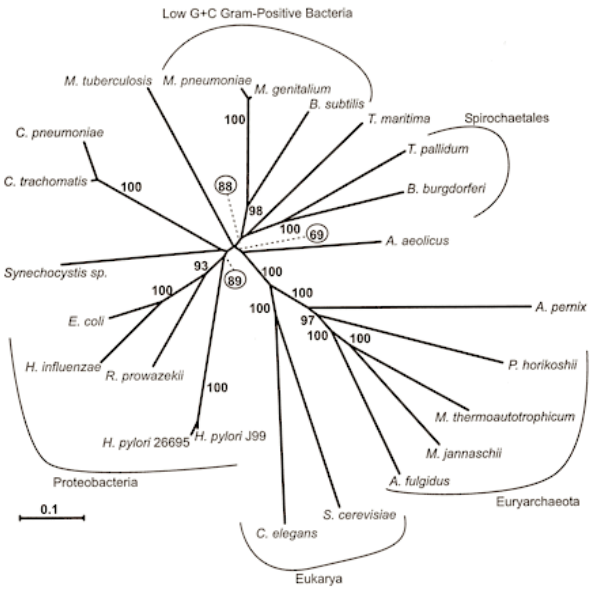


Fig. 2.4: Arbre basé sur la fraction de gènes orthologues partagée entre génomes. L'arbre est construit par la méthode de Neighbor joining sur des distances définies comme le nombre de gènes communs à deux bactéries, pondéré par le nombre de gènes contenus dans le plus petit des deux génomes.
Extrait de Huynen, *et al.*, 1999, une mise à jour de Snel *et al.* (1999)

L'organisation des gènes sur les chromosomes a été utilisée pour reconstruire la phylogénie des bactéries (Snel, *et al.*, 1999; Wolf, *et al.*, 2001), mais permet de ne retrouver que les liens de parenté entre espèces proches, comme il était prévisible au regard de la fig. 2.2. L'une des premières méthodes proposant d'utiliser le contenu en gène des génomes pour inférer les relations de parenté entre les organismes est le travail

de Snel *et al.* (Snel, *et al.*, 1999). La similarité entre deux génomes est définie par le rapport du nombre de gènes orthologues qu'ils ont en commun et du nombre de gènes du plus petit des deux génomes. C'est une définition opérationnelle, et non phylogénétique, de l'orthologie qui est utilisée ici : deux gènes sont considérés comme étant orthologues si, en utilisant l'algorithme de recherche de similarité de Smith et Waterman (Smith et Waterman, 1981), les deux gènes sont réciproquement détectés comme étant les plus similaires dans les deux génomes considérés. La phylogénie obtenue par Snel *et al.* (Snel, *et al.*, 1999) permet de retrouver la plupart des groupes proposés par celle basée sur l'ARN ribosomal, notamment celui des protéobactéries et des bactéries gram-positives à bas G+C (Fig. 2.4). Bien que certains branchements profonds restent irrésolus, la congruence de ces deux arbres suggère deux choses : d'une part, que les parties résolues de l'arbre de l'ARN ribosomal représentent bien une réalité phylogénétique, et d'autre part que l'abondance des acquisitions et des pertes de gènes au cours de l'évolution n'est pas suffisante pour brouiller complètement le signal phylogénétique que représente le partage de gènes orthologues entre espèces. Cependant, des interprétations différentes de ces résultat ont été suggérées qui proposent que ces arbres basés sur le contenu en gènes ne reflètent que la propension des espèces à échanger des gènes par transfert horizontal (Doolittle, 1999b).

D'autres méthodes proposent de classer les espèces sur la base non pas des gènes orthologues, mais de la présence de familles de gènes, ce qui est sensiblement différent puisque aucune différence n'est faite entre les classes d'homologies (orthologie/paralogie). Le fait de considérer comme un caractère la présence d'une famille de gènes conduit d'une part à réduire la quantité d'information phylogénétique (une famille de gènes constitue au plus un caractère informatif alors qu'elle peut contenir plusieurs familles d'orthologues) et d'autre part à réduire sa qualité (la prise en compte des gènes paralogues bruite le signal phylogénétique). Ces méthodes donnent des résultats variables même avec des protocoles

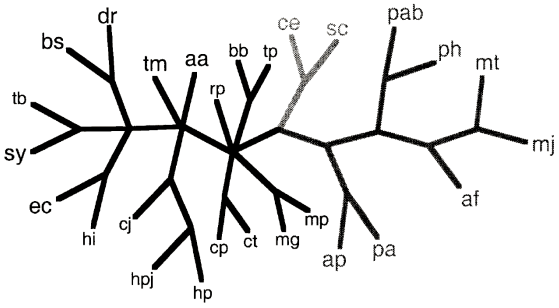


Fig 2.5 : Arbre basé sur la présence/absence de familles de gènes. On identifie le groupe de archées à droite, les eucaryotes au milieu (en clair) et les bactéries à gauche. A la base des bactéries on trouve les espèces dont le génome est réduit comme Rickettsia (rp) (loin des protéobactéries comme *E. coli* – ec), les mycoplasmes (mp et mg) (loin des gram-positives à bas G+C comme *B. subtilis* – bs), les chlamydiales (cp et ct) et les spirochètes (tp et bb). Extrait de House et Fitz-Gibbon, 2002.

très semblables (Fitz-Gibbon et House, 1999; Tekaia, *et al.*, 1999; Lin et Gerstein, 2000; House et Fitz-Gibbon, 2002). D'une manière générale, elles ne réussissent qu'à retrouver les trois grands domaines de la vie (eucaryotes, bactéries et archées), ou les groupes d'espèces très proches. De plus, ces méthodes semblent être extrêmement sensibles à la taille des génomes. Ceci suggère l'existence de familles domaine-spécifiques (ou n'ayant pu être regroupées du fait de la forte divergence entre ces domaines) et que des familles entières de gènes sont absentes des génomes ayant subi une réduction de leur taille. Par exemple dans la Fig. 2.5 l'arbre construit par House et Fitz-Gibbon (House et Fitz-Gibbon, 2002) regroupe à la base des bactéries les espèces dont les génomes sont les plus réduits comme les mycoplasmes, les chlamydiales, *Rickettsia prowazekii* et les spirochètes.

2.1.2.2 Prise en compte de la similarité des séquences

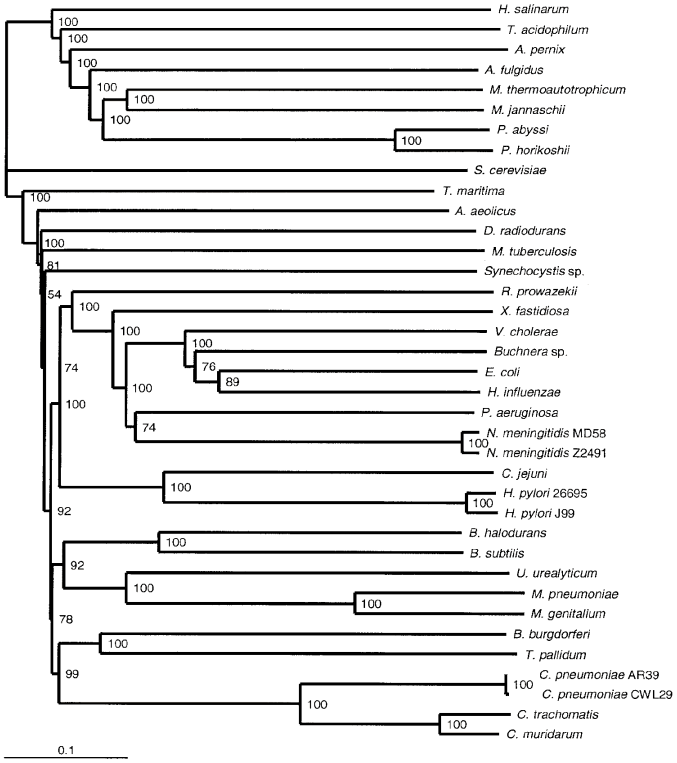


Fig. 2.6 : Arbre basé sur la moyenne des scores de BLASTP normalisés. La topologie de l'arbre montre une congruence remarquable avec la phylogénie de l'ARN ribosomal.
Extrait de Clarke, *et al.*, 2002.

Même si certaines méthodes basées sur le contenu en gène (et notamment sur les orthologues partagés) permettent de retrouver des groupes congruents avec l'ARN ribosomal, on peut regretter le fait que toutes ignorent complètement l'information que contiennent les séquences. En effet, la figure 2.2 montre que le pourcentage de similarité entre protéines constitue un meilleur indicateur de l'évolution des génomes, notamment pour de grandes distances évolutives. Pour intégrer ces données, d'autres approches ont été proposées qui prennent en compte la similarité des gènes dans les couples de génomes. Grishin *et al.* (Grishin, *et al.*,

2000) proposent ainsi d'estimer les distances entre espèces en prenant en compte la similarité des séquences prises deux à deux. L'arbre obtenu soutient une position basale de la bactérie hyperthermophile *Aquifex aeolicus* et de la cyanobactérie *Synechocystis*, mais échoue à résoudre le reste de l'arbre des bactéries. D'autres études plus récentes (Wolf, *et al.*, 2001; Clarke, *et al.*, 2002) utilisent le même type d'information avec plus de succès. Notamment, Clarke *et al.* (Clarke, *et al.*, 2002) décrivent une méthode d'estimation des distances entre génomes basée sur les scores de BLASTP et proposent un arbre assez robuste qui présente de grandes similitudes avec l'arbre basé sur l'ARN ribosomal. Cependant, les indices de robustesse des noeuds que proposent ces auteurs sont difficiles à interpréter car il s'agit d'échantillonner des sous-ensembles de gènes partagés par deux espèces pour calculer une distance. Contrairement à ce que supposent ces auteurs, ce type de ré-échantillonnage est assez éloigné du principe du bootstrap en phylogénie, et le soutien statistique des nœuds de leur arbre peut donc être mis en doute. Le rapport des longueurs des branches internes et terminales laisse au contraire supposer que cet arbre n'est en fait pas résolu.

2.1.2.3 Remarques sur la définition d'orthologie

Plusieurs remarques peuvent être faite à propos des méthodes liées au contenu en gènes, notamment si l'on considère que toutes utilisent la définition citée plus haut de meilleure similarité réciproque. Cette définition est largement utilisée par les biologistes bien qu'elle puisse potentiellement conduire à des erreurs relativement importantes. L'un des problèmes majeurs lié à la prise en considération de seulement deux organismes pour déterminer les relations d'orthologie est celui des paralogies cachées. Par exemple, dans la famille de la Glutamate synthase, une recherche entre *Bacillus subtilis* et *Synechocystis sp.* devrait donner comme orthologues les gènes GLTB_BACSU et GLTB_SYNY3 ce qui semble correct au regard de la phylogénie (Fig. 2.7). Cependant, entre *E. coli* et *B. subtilis*, la recherche d'orthologues devrait proposer les gènes GLTB_ECOLI et GLTB_BACSU. Or, dans la phylogénie de la famille, la présence de deux gènes de Glutamate synthase chez *Synechocystis* ainsi que chez *Vibrio cholerae* atteste de l'existence d'une probable duplication antérieure à l'ancêtre commun de toutes ces espèces dans cette famille et révèle que les deux gènes sont en fait des paralogues. Il semble, au regard de la phylogénie, que *Bacillus* a perdu une des copies de ce gène tandis que *Escherichia* perdait l'autre copie. On comprend aisément

que ce type d'erreur est d'autant plus important que le nombre de génomes considéré est faible.

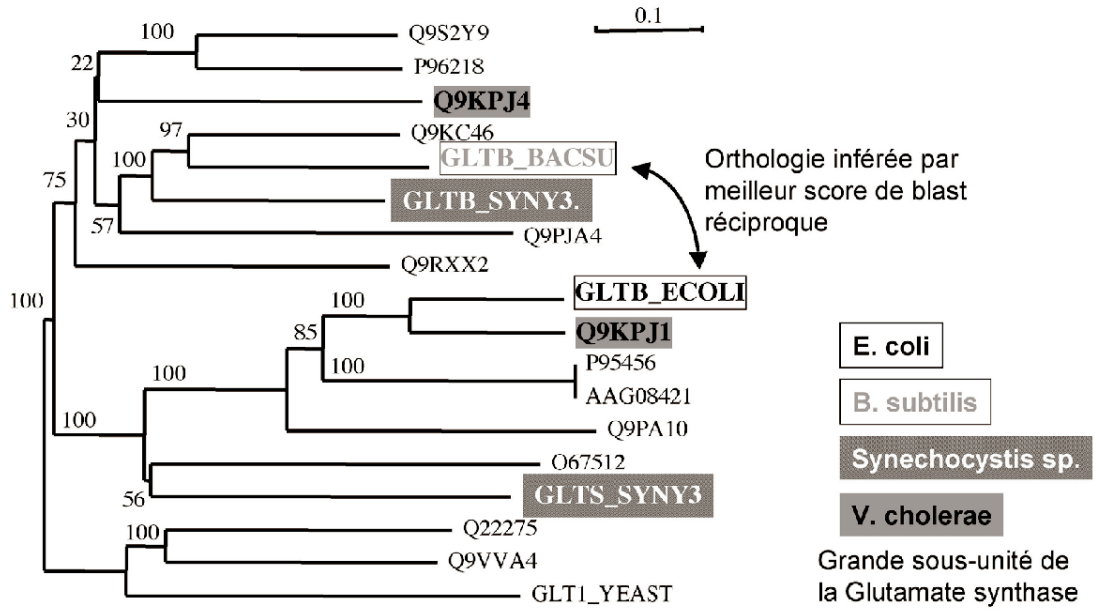


Fig. 2.7 : Arbre construit à partir de la grande sous unité de la Glutamate synthase. Voir détails dans le texte

Un autre problème lié à la définition d'orthologie par similarité est dû au fait qu'il n'existe pas une relation simple entre la similarité de deux séquences et leur proximité phylogénétique. En effet, Koski et Golding (Koski et Golding, 2001) ont montré que les séquences les plus proches phylogénétiquement pouvaient très souvent ne pas correspondre à la meilleure similarité détectée par BLAST (Altschul, *et al.*, 1997), et ceci d'autant plus que les organismes considérés sont éloignés. Par exemple, ils montrent que pour *Aeropyrum pernix*, une archée pour laquelle peu de séquences d'organismes proches étaient disponibles à l'époque, plus de 40 % des meilleurs scores de BLAST n'étaient pas les plus proches voisins phylogénétiquement, et 30 % n'étaient même pas dans le même domaine de la vie. Ce chiffre tombe à 27 % (7 % se trouvant dans un autre domaine de la vie) lorsque c'est *E. coli*, dont de nombreuses espèces proches sont séquencées, qui est considérée, ce qui reste malgré tout extrêmement élevé. Le critère de réciprocité doit permettre de corriger un certain nombre de ces aberrations, mais il reste à déterminer dans quelle mesure.

2.1.2.4 Autres mesures de distance proposées

On peut citer également d'autres indices de ressemblance entre les génomes, comme la composition en dinucléotides, en acide aminés ou en motifs structuraux des protéines, qui ont donné lieu à des tentatives de représentation phylogénétique plus ou moins probantes (voir par exemple Lin et Gerstein, 2000). Cependant, bien que certains auteurs suggèrent que la composition en dinucléotides des génomes puisse constituer un marqueur phylogénétique (Karlin, *et al.*, 1997 ; Brocchieri, 2001), ceci essentiellement sur la base d'études entre espèces proches (entre espèces du même genre) ou très éloignées (entre bactéries et archées par exemple), son application à des problèmes phylogénétiques moins triviaux reste très hasardeuse. En ce qui concerne la composition en acides aminés des protéines, elle a été montrée comme étant fortement dépendante des conditions de vie des organismes, notamment chez les hyperthermophiles (Kreil et Ouzounis, 2001), ce qui en fait un très mauvais caractère phylogénétique. Enfin, les mesures basées sur le partage de certains motifs structuraux dans les protéines se heurtent aux mêmes problèmes que les méthodes basées sur la présence/absence de familles multi-géniques (Lin et Gerstein, 2000).

Ainsi, il semble souhaitable pour résoudre un problème tel que celui de la phylogénie des bactéries, d'être capable à la fois de prendre en compte l'information phylogénétique que contiennent les séquences, de ne pas se limiter aux rares gènes présents chez tous les organismes considérés, et surtout de pouvoir limiter l'impact des familles de gènes sujettes à transferts horizontaux. Ceci nous conduit à considérer les méthodes permettant d'identifier des informations congruentes entre les gènes.

2.2 Les tests de congruence entre les données phylogénétiques

Le problème d'évaluer si des gènes peuvent contenir des informations congruentes, et de l'identification de ces gènes est abordé dans de nombreux articles avec des méthodes diverses. (Rivera, *et al.*, 1998; Jain, *et al.*, 1999; Nesbo, *et al.*, 2001; Brochier, *et al.*, 2002; Matte-Tailliez, *et al.*, 2002; Zhaxybayeva et Gogarten, 2002). Nous avons déjà parlé d'une méthode empirique, appliquée notamment par Teichmann et Mitchison (Teichmann et

Mitchison, 1999) et Brown *et al.* (Brown, *et al.*, 2001) qui consiste à considérer les topologies obtenues pour chaque gène, et à les comparer à une référence (en l'occurrence la phylogénie de l'ARN ribosomal). Comme aucun des arbres n'est strictement identique à la référence, les auteurs choisissent alors comme critère d'exclusion des familles, les différences qu'ils jugent importantes (notamment, la non monophylie des domaines). Nous avons vu que cette approche était susceptible d'augmenter l'importance de certains regroupement illégitimes par rapport à d'autres (section 2.1). D'autres méthodes moins subjectives ont été proposées.

2.2.1 ***Comparaison topologique***

Jain *et al.* (Jain, *et al.*, 1999) utilisent une approche topologique pour identifier les transferts horizontaux de 312 familles de gènes orthologues présents chez six espèces (quatre bactéries et deux archées). En utilisant un indice de distance topologique défini comme le nombre de nœuds qu'une branche terminale doit traverser pour réconcilier l'arbre analysé et la référence (en l'occurrence, une phylogénie basée sur le facteur d'élongation EF-1α), ils montrent que de très nombreux transferts de gènes se sont produits dans les arbres analysés. Ils notent que ces transferts se sont produits depuis la diversification des groupes de bactéries représentés dans l'arbre et que les gènes impliqués dans des fonctions liées à la gestion de l'information génétique (réplication, transcription et traduction) sont nettement moins affectés que les autres. Le but de cette étude n'est pas de reconstruire les liens de parenté entre les espèces puisqu'ils sont considérés comme étant connus *a priori*. Ce point est regrettable car la topologie considérée ici comme référence, est loin d'être consensuelle. De plus, les incongruences sont interprétées en terme de transferts horizontaux uniquement. Or, pour une phylogénie basée sur six espèces très éloignées, avec une identification des orthologues par la méthode de double meilleurs scores réciproques, les risques de paralogies cachées sont loin d'être négligeables. De plus, même pour une famille de vrais orthologues,

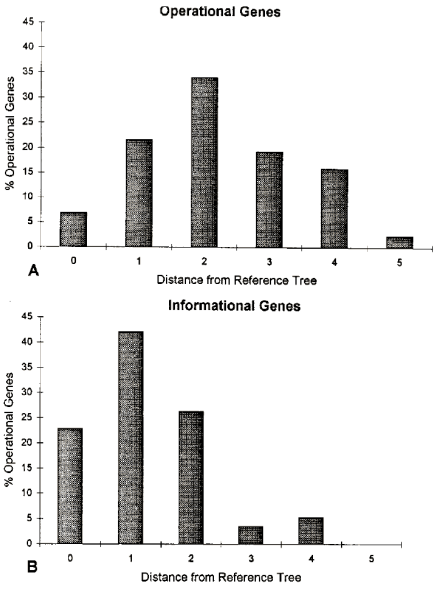


Fig. 2.8 : Distribution des distances topologiques des arbres construits à partir de différentes familles à un arbre de référence (basé sur le facteur d'élongation EF-1α). Extrait de Jain, *et al.*, 1999. Voir détails dans le texte

les méthodes de reconstruction ne garantissent absolument pas d'obtenir la topologie vraie. Ainsi, outre l'interprétation des résultats par l'hypothèse des transferts horizontaux, on peut suggérer soit que les gènes opérationnels possèdent plus fréquemment des paralogues cachés, soit qu'ils sont moins contraints par la sélection, évoluent plus vite et contiennent moins de signal phylogénétique, soit un mélange des trois hypothèses.

2.2.2 ***Likelihood mapping***

Une méthode élégante de comparaison des alignements à quatre taxons a été proposée par Strimmer et von Haesler (Strimmer et von Haeseler, 1997). Appelée Likelihood- (ou Quartet-) Mapping, elle consiste à évaluer la probabilité que l'alignement ait été produit par chacun des trois quartets (arbres contenant quatre taxons) possibles. En utilisant un système de coordonnées barycentriques, ce résultat est ensuite représenté par un point dans un triangle isocèle, où la distance du point aux trois arêtes du triangle représente les probabilités de chacun des trois arbres (Fig. 2.9). Un alignement contient une information d'autant plus explicite sur la phylogénie des quatre espèces que le point qui le représente est proche d'un sommet. Ainsi, on peut également considérer que chaque sommet du triangle représente un arbre et que plus le point de coordonnées (p_1, p_2, p_3) est proche d'un des sommets, plus l'arbre correspondant est probable. Ce type de représentation permet d'évaluer de nombreux alignements en même temps, et de voir si l'un des trois sommets rassemble plus de points que les deux autres. Cette méthode a été adaptée à l'étude des transferts horizontaux chez les procaryotes dans au moins deux études (Nesbo, *et al.*, 2001; Zhaxybayeva et Gogarten, 2002). Toutes deux concluent à l'impossibilité d'identifier un cœur de gènes ayant partagé une histoire commune du fait de trop nombreux transferts horizontaux. Cependant, Nesbo *et al.* (Nesbo, *et al.*, 2001) admettent que ces résultats peuvent également s'interpréter comme une perte du signal phylogénétique, ce qui ne rend pas leur

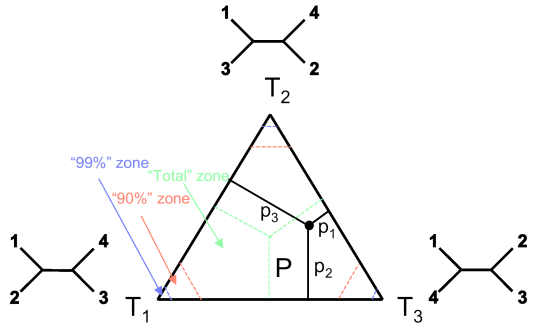


Fig. 2.9 : Le Likelihood Mapping. La probabilité des trois topologies possibles est représentée par les distances p_1, p_2 et p_3. Si le point est dans le coin représenté par la topologie T_1, cela signifie que la probabilité de l'arbre T_1 est très supérieure aux deux autres. Extrait de Zhaxybayeva et Gogarten, 2002

conclusion moins pessimiste quant à l'identification de gènes informatifs sur les relations phylogénétiques entre espèces. Il est néanmoins important de noter que les résultats basés sur la reconstruction d'arbres à quatre espèces doivent être pris avec beaucoup de précautions. Phillippe et Douzeri (Philippe et Douzery, 1994) pour la parcimonie et plus tard Adachi et Hasegawa (Adachi et Hasegawa, 1996) pour le maximum de vraisemblance, ont en effet montré que les phylogénies basées sur un très faible nombre d'espèces sont peu fiables et que « reconstruire l'histoire avec seulement quatre taxons est plutôt un jeu de hasard » (Philippe et Douzery, 1994). Ceci correspond assez bien aux résultats de Zhaxybayeva et Gogarten (Zhaxybayeva et Gogarten, 2002) puisque pour de nombreux quartets d'espèces, chaque arbre est soutenu par une même proportion d'alignements. D'autre part, Strimmer et von Haesler (Strimmer et von Haeseler, 1997) puis Nieselt-Struwe et von Haesler (Nieselt-Struwe et von Haeseler, 2001), étudiant les propriétés de leur méthode, ont mis en garde contre un certain nombre de facteurs influant sur les résultats et qui n'ont pas été pris en compte dans les analyses des génomes procaryotes, comme la sensibilité de la méthode à la longueur des alignements, et au modèle d'évolution utilisé.

2.2.3 *ACP sur les valeurs de vraisemblance*

2.2.3 *ACP sur les valeurs de vraisemblance*

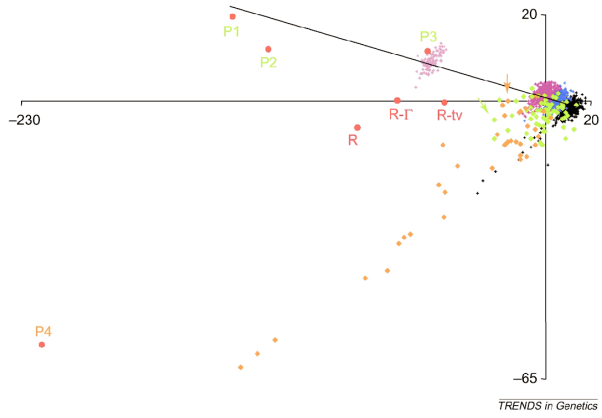


Fig. 2.10: ACP sur les valeurs de vraisemblance de différentes familles protéiques et nucléiques pour 375 topologies d'arbres des bactéries. Les points représentent les alignements dans l'espace des topologies. La plupart des alignements soutiennent des arbres similaires et se regroupent au niveau de l'origine du graphe. Les points P1, P2, P3 et P4 représentent différents concaténats de protéines. Les points R représentent l'alignement de l'ARN ribosomal en utilisant différentes méthodes de distance. Voir détails dans le texte. Extrait de Brochier, *et al.*, 2002.

Pour résoudre le problème notamment de l'utilisation de quartets, Brochier *et al.* (Brochier, *et al.*, 2002) ont proposé une méthode particulièrement intéressante : grâce à l'utilisation d'une méthode d'analyse multivariée, l'analyse en composante principale (ACP), ils proposent une représentation graphique de la congruence des alignements. Un certain nombre de topologies concurrentes sont analysées par la méthode de maximum de vraisemblance pour chacun 57 alignements considérés (les familles de gènes de la machinerie de traduction). Il en résulte un tableau contenant les valeurs de

vraisemblance de chaque alignement (en lignes) pour chacune des topologies (en colonnes). L'ACP permet une représentation de ce tableau dans les dimensions qui en maximisent la variance. Ainsi, les alignements soutenant des arbres proches sont regroupés graphiquement.

L'une des difficultés principales de cette approche est le choix des topologies. En effet, l'échantillonnage taxonomique considéré dans cette étude est de 45 espèces, ce qui correspond à un nombre astronomique de topologies possibles (près de 10^{64}). Un choix doit donc s'opérer sur les topologies analysées. Les auteurs proposent un choix raisonnable qui est de prendre les meilleures topologies pour chacun des alignements, et identifient ainsi un nombre important de gènes portant une information phylogénétique congruente. Ils confirment ainsi l'existence d'un ensemble de gènes ayant connu des histoires parallèles au cours de l'évolution. À partir de plusieurs dizaines de ces gènes concaténés, ils infèrent une phylogénie des bactéries relativement proche de celle de l'ARN ribosomal, mais où certaines relations entre les grandes divisions bactériennes apparaissent plus clairement (nous reviendrons sur ce travail section 2.3.3.2). Ce faisant, ils concatènent des gènes qui ne sont pas présents chez toutes les espèces considérées, mais notent que le traitement adéquat d'un « super-alignement » obtenu à partir de plusieurs dizaines de gènes nécessite d'utiliser une méthode de reconstruction phylogénétique capable de prendre en compte la grande variabilité des modalités d'évolution des différents gènes. De telles méthodes, bien qu'abordées sur le plan théorique ne sont pas encore disponibles. Une autre limite de cette méthode est, comme nous l'avons déjà noté (voir section 2.1.1) que les gènes utilisés doivent être représentés dans un nombre important d'espèces (Brochier*, et al.*, 2002 n'ont considéré que les gènes présents chez au moins 45 espèces représentant raisonnablement la diversité des bactéries).

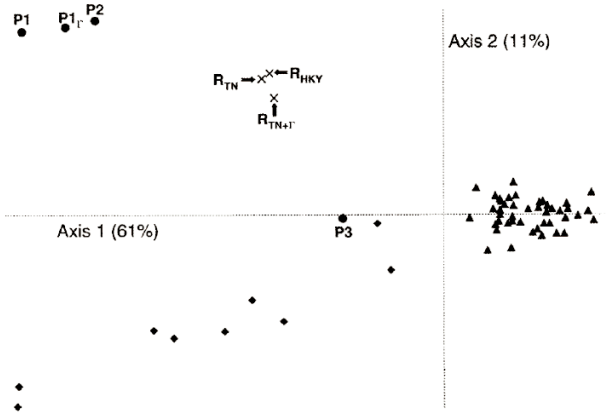


Fig. 2.11 : ACP sur les valeurs de vraisemblance de 53 familles protéiques et nucléiques pour 49 topologies d'arbres des archées. Les familles n'ayant probablement pas subi de transferts sont regroupés à droite. Lorsque ces familles sont concaténées, elles donnent l'alignement P2. Les autres points représentent des familles dans lesquelles des transferts horizontaux ont eu lieu. Ces dernières, lorsqu'elles sont concaténées donnent l'alignement noté P3. Voir détails dans le texte. Extrait de Matte-Tailliez*, et al.*, 2002.

La même méthode a été appliquée à l'étude de 53 protéines ribosomales de 14 espèces d'archées (Matte-Tailliez, *et al.*, 2002). Huit gènes ayant probablement subi des transferts horizontaux sont ainsi identifiés. De manière surprenante, les huit gènes concaténés soutiennent une phylogénie très similaire à celle qui maximise la vraisemblance des 45 protéines n'ayant pas subi de transfert (ce concaténat est représenté par le point P3 sur la fig. 2.11). D'autre part, la phylogénie basée sur la concaténation de ces 45 protéines est très différente des arbres qui sont responsables du regroupement des gènes dans l'ACP (notez l'éloignement du point P2 et du nuage de point des 45 gènes sur les deux axes fig. 2.11). Le même phénomène, quoique moins prononcé avait déjà été observé dans l'étude sur les gènes bactériens (Brochier, *et al.*, 2002) ce qui suggère que les méthodes actuellement disponibles pour traiter les jeux de séquences concaténées sont incapables de prendre en compte la diversité des modalités d'évolution des différents gènes.

2.3 Une approche topologique : le superarbre

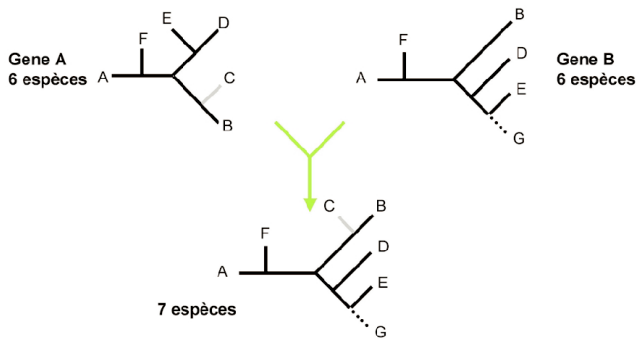


Fig. 2.12 : principe du superarbre. A partir de deux arbres ayant un échantillonnage taxonomique différent mais recouvrant, il est possible d'inférer certaines des relations entre des espèces qui sont absentes dans l'un des arbres.

Ainsi, plusieurs problèmes se posent encore, notamment pour la prise en compte des gènes cantonnés à certains phylum bactériens, portant une information phylogénétique sur des sous parties de l'arbre et non sur tout l'arbre. Une solution possible à ce problème est la combinaison non pas de séquences, mais d'arbres. L'approche de reconstruction de super-arbres permet, comme l'indique la fig. 2.12, d'inférer par exemple un arbre à sept espèces à partir de deux arbres à six espèces dont l'échantillonnage taxonomique se recouvre. Cette approche autorise ainsi la combinaison de grandes quantités de données concernant un vaste échantillonnage taxonomique.

La première partie de ce chapitre concerne l'application de cette méthode aux données génomiques pour l'inférence d'une phylogénie racinée des bactéries.

2.3.1 ***Matériels et méthodes***

Nous ne présentons ici que la dernière version du travail dédié à la phylogénie des bactéries inférée par la méthode de super-arbre, et qui a donné lieu a une publication dans *Genome Research* (Daubin, *et al.*, 2002). Il sera pourtant fait allusion aux versions précédentes de ce travail (présentées à Jobim 2001 et dans Daubin, *et al.*, 2001) dans la mesure où elles permettent d'expliquer des choix méthodologiques de la présente approche. Nous insistons sur le fait que le travail présenté ici vise principalement à reconstruire une phylogénie racinée des bactéries et non à reconstruire la phylogénie universelle. La nuance réside essentiellement dans la moindre attention portée aux familles de gènes n'ayant aucun représentant bactérien, et notamment les familles de gènes exclusivement eucaryotes qui n'ont pas été prises en compte dans cette étude.

2.3.1.1 Construction des familles de gènes : HOBACGEN-CG.

HOBACGEN (HOmologous BACterial GENes, Perriere, *et al.*, 2000b) est une banque de séquences qui regroupe en familles homologues les gènes de l'ensemble des organismes procaryotes et de la levure. Nous avons utilisé la procédure de construction d'HOBACGEN pour développer HOBACGEN-CG (pour « Complete Genome »), une banque qui regroupe l'ensemble des espèces dont le génome a été complètement séquencé. La deuxième version de cette banque contenait 45 espèces, dont 32 bactéries, neuf archées et quatre eucaryotes (Tableau 2.1).

La première étape de construction de la banque a été de récupérer toutes les séquences présentes dans les banques protéiques SWISS-PROT et TrEMBL pour ces 45 espèces. On effectue ensuite une recherche de similarité de toutes les séquences contre elles-mêmes en utilisant le programme BLASTP2 (Altschul, *et al.*, 1997). Les séquences possédant plus de 50% de similarité (BLOSUM62) sur plus de 80 % de leur longueur sont intégrées dans la même famille. Une relation d'inclusion par simple lien est ajoutée, si bien qu'il suffit à une séquence de remplir ce critère pour une seule des séquences de la famille pour en faire partie. Les séquences protéiques ainsi sélectionnées sont ensuite alignées à l'aide de CLUSTALW

(Higgins, *et al.*, 1996) et un arbre phylogénétique est reconstruit en utilisant le programme BIONJ (Gascuel, 1997).

Les arbres, qui constituent surtout un moyen pratique de visualiser la famille, peuvent ensuite être consultés via l'interface FamFetch (Perriere, *et al.*, 2000b) qui permet en outre de faire de nombreuses requêtes sur les familles et les arbres.

Bactéries (32 espèces)	
Protéobactéries (12 espèces)	γ€: *Escherichia coli, Vibrio cholerae, Pasteurella multocida, Haemophilus influenzae, Buchnera sp., Pseudomonas aeruginosa, Xylella fastidiosa,* β€: *Neisseria meningitidis,* α€: *Caulobacter crescentus, Rickettsia prowazekii,* ε€: *Helicobacter pylori, Campylobacter jejuni.*
Gram-positive Bas-G+C (8 espèces)	*Bacillus subtilis, Bacillus halodurans, Staphylococcus aureus, Lactococcus lactis, Streptcoccus pyogenes, Mycoplasma pneumoniae, Mycoplasma genitalium, Ureaplasma parvum.*
Gram-positive Haut-G+C (3 espèces)	*Mycobacterium tuberculosis, Mycobacterium leprae, Streptomyces coelicolor.*
Cyanobactérie (1 espèce)	*Synechocystis sp.* (PCC 6803)
Spirochètes (2 espèces)	*Borrelia burgdorferi, Treponema pallidum.*
Chlamydiales (3 espèces)	*Chlamydia muridarum, Chlamydia trachomatis, Chlamydophila pneumoniae.*
Deinococcus/Thermus (1 espèce)	*Deinococcus radiodurans.*
Hyperthermophiles (2 espèces)	*Aquifex aeolicus, Thermotoga maritima.*
Archées (9 espèces)	
Euryarchaeotes (7 espèces)	*Halobacterium sp., Thermoplasma acidophilum, Methanococcus jannashii, Pyrococcus horikoshii, Pyrococcus abyssi, Methanobacterium thermoautotrophicum, Archaeoglobus fulgidus.*
Crenarchaeotes (2 espèces)	*Sulfolobus solfataricus, Aeropyrum pernix.*
Eucaryotes (4 espèces)	*Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae.*

Tableau 2.1 : Les espèces représentées dans HOBACGEN-CG release 2 et présentes dans le superarbre.

2.3.1.2 Première sélection des familles

La reconstruction de phylogénies d'espèces à partir de données moléculaires nécessite de prendre quelques précautions. En effet, comme nous l'avons vu précédemment, plusieurs types d'homologues peuvent être identifiés sur la base de la similarité des séquences, mais seule une classe d'homologie, les orthologues, permettent l'étude des relations entre espèces. La meilleure manière d'identifier des gènes orthologues est de prendre tous les homologues disponibles et de reconstruire une phylogénie. C'est cette approche que nous avons choisi. Cependant, même ainsi, certaines relations sont peu claires si l'on se refuse à faire des hypothèses *a priori* sur la topologie du véritable arbre des bactéries. Dans ces conditions, toute famille contenant plusieurs gènes de la même espèce, s'ils ne forment pas un groupe monophylétique, est susceptible de contenir des paralogies cachées, non seulement en ce qui concerne cette espèce mais également toutes les autres. En effet, si l'on considère la phylogénie de la Glutamate synthase (Fig. 2.7), en l'absence des séquences de *Vibrio* et *Synechocystis*, on peut considérer que les gènes présents sont tous orthologues. Cependant, la prise en compte de ces séquences permet de mettre en évidence la possibilité que cette famille ne soit pas constituée uniquement de gènes orthologues. Deux hypothèses peuvent en effet expliquer la topologie observée : des transferts horizontaux n'ayant concerné que *Vibrio* et *Synechocystis* (dans ce cas, il suffirait de ne pas considérer ces séquences pour avoir une famille représentant la phylogénie des espèces), ou bien une duplication ancestrale suivie de pertes différentielles dans les différentes espèces. Le fait de trancher entre ces deux hypothèses constitue un gros risque, car il est très difficile si l'on ne fait pas d'hypothèse *a priori* sur la phylogénie, de considérer que l'une ou l'autre est plus parcimonieuse. Ainsi, la présence des séquences de *Synechocystis* et *Vibrio* interdit de considérer les gènes de *Bacillus* et *Escherichia* comme des orthologues fiables. Nous avons choisi une méthode très stringente qui consiste à exclure dans ce cas la famille de l'analyse.

Dans un souci de minimiser l'impact des transferts horizontaux, nous avons également exclu des familles tous les gènes eucaryotes dont le produit est connu pour avoir une localisation mitochondriale ou chloroplastique. De même, les familles ne contenant que des séquences d'archées et de bactéries hyperthermophiles n'ont pas été retenues du fait de la forte présomption de transfert qui pèse sur ces gènes. D'une manière plus générale, les familles ne contenant qu'une ou deux séquences bactériennes ont également été exclues.

2.3.1.3 Reconstruction des arbres

Chaque famille ainsi sélectionnée a ensuite été réalignée en utilisant le programme CLUSTALW, et les parties fiables des alignements ont été sélectionnées en utilisant le programme Gblocks (Castresana, 2000). Seuls les alignement pour lesquels ce traitement laissait au moins deux fois plus de sites que de séquences ont été retenus. A partir de ces alignements, deux analyses phylogénétiques ont été conduites indépendamment : une analyse par maximum de vraisemblance avec le programme PROTML (Kishino, *et al.*, 1990) et le modèle de substitution JTT (Jones, *et al.*, 1992) ; et une analyse par le programme BIONJ (Gascuel, 1997) en utilisant une distance calculée grâce à une loi Gamma implémentée dans TREE-PUZZLE (Strimmer et von Haeseler, 1996) et le même modèle de substitution que précédemment: JTT (Jones, *et al.*, 1992).

Ces deux méthodes permettent d'avoir des indices de confiance aux nœuds : RELL-BP pour PROTML et le *bootstrap* pour BIONJ.

2.3.1.4 Deuxième sélection des familles.

Le souci premier de cette analyse était de ne faire aucune hypothèse *a priori* sur la topologie de l'arbre des bactéries. Cependant, les résultats des analyses précédentes (Daubin, *et al.*, 2001) présentaient un ressemblance frappante avec ceux de Brown, *et al.*, 2001 (Fig 2.1A), notamment en ce qui concerne la position de la racine du domaine des bactéries dans la branche des spirochètes. Ces auteurs interprètent ce résultat comme étant dû a des transferts inter-domaines (voir section 2.1.1). Afin d'exclure l'éventualité d'un tel artefact dans notre analyse, nous avons ajouté un critère de sélection des familles pour le présent travail : la monophylie du domaine des bactéries. Ce domaine est très largement considéré comme étant monophylétique. La mise en oeuvre de ce critère nous a conduit à exclure quelques familles et à supprimer certaines séquences dont la position dans les arbres suggérait un transfert évident (dont notamment des gènes eucaryotes d'origine mitochondriale non annotés comme tels).

2.3.1.5 Méthode de Représentation de Matrice par Parcimonie (MRP)

La méthode de construction de super-arbre choisie pour cette étude est celle de la Représentation de Matrice par Parcimonie ou MRP (Baum, 1992; Ragan, 1992). Cette méthode avait à l'origine été proposée afin de combiner des jeux de données de nature différente (comme des données morphologiques et des données moléculaires). Elle a plus récemment été proposée pour combiner des arbres de la littérature et inférer une phylogénie des placentaires (Liu, *et al.*, 2001). Son principe, ainsi que les adaptations que nous y avons apportés sont décrits dans la Fig. 2.13. Chaque nœud d'un arbre est recodé dans une matrice binaire par un caractère qui représente la bipartition correspondante. Nous avons introduit un seuil au codage des nœuds et seuls ceux soutenus par des valeurs de bootstrap ou de RELL-BP supérieures à 50 % sont codés. Les arbres obtenus par les deux méthodes sont traités indépendamment.

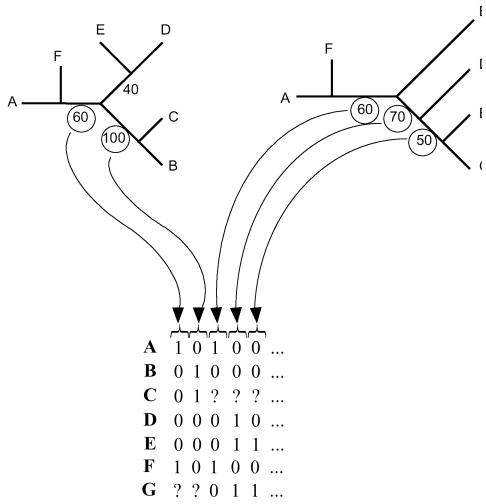


Fig. 2.13 : La méthode de représentation en matrice de parcimonie (MRP). Chaque branche interne est codée par un caractère binaire qui décrit la bipartition correspondante. Seuls les nœuds ayant une valeur de bootstrap supérieure à 50 % ont été codés. Les matrices correspondant à chaque arbre peuvent être concaténées en ajoutant des caractères manquants « ? ».

Les matrices obtenues après codage des arbres sont traitées par la méthode de parcimonie en faisant à chaque fois 500 réplicats de bootstrap. Pour cela, les programmes du package PHYLIP (Felsenstein, 1989) ont été utilisé.

2.3.1.6 Comparaison entre arbres

Certains arbres particulièrement aberrant au niveau de leur topologie peuvent introduire du bruit dans l'analyse et fausser la reconstruction du super-arbre. Il nous faut donc un critère pour éliminer ces arbres. La méthode MRP étant basée sur les topologies des arbres, il est donc judicieux de sélectionner les arbres sur un critère de similarité topologique. Nous

avons pris le parti dans cette étude de ne considérer aucune topologie comme bonne *a priori*. Une solution pour éviter de comparer tous les arbres à une topologie de référence est de faire toutes les comparaisons possibles entre arbres et de visualiser les résultats par une analyse multivariée.

La première étape est de réduire les deux arbres aux taxons qu'ils ont en commun Nous avons choisi une distance analogue à la distance de Robinson et Foulds (Robinson et Foulds, 1981), qui est définie comme le nombre minimal de nœuds qu'il faut faire traverser à des branches pour transformer une topologie en une autre. Cette mesure de distance topologique revient à dénombrer les bipartitions communes aux deux arbres. Notre indice de distance se définit donc ainsi :

$$D = 1 - \frac{b_c}{b_t}$$

Où b_c est le nombre de bipartitions communes aux deux arbres et b_t est le nombre total de bipartitions. Cette distance varie donc de 0 pour deux arbres identiques, à 1 pour deux arbres ne possédant aucune bipartition commune.

Toutes les distances entre couples d'arbres possédant un nombre minimum d'espèces en commun peuvent ainsi être calculées.

2.3.1.7 L'Analyse en Coordonnées Principales ou ACO (PCO en anglais).

Si n est le nombre d'arbres, on obtient donc une matrice de distance symétrique de dimension $n \times n$. Une difficulté est que les arbres n'ont pas toujours suffisamment d'espèces en commun pour qu'une distance puisse être calculée. Pour minimiser le nombre de trous ainsi occasionnés dans la matrice, nous avons décidé de réduire l'analyse de la congruence entre les arbres à la partie bactérienne des arbres, en ne sélectionnant pour l'ACO que ceux contenant au moins 10 espèces de bactéries. Même ainsi, certains arbres restent impossibles à comparer, et dans ce cas, le meilleur estimateur des données manquantes de la matrice est la moyenne de toutes les distances présentes dans la matrice (D. Chessel, communication personnelle). La matrice peut-être analysée par la méthode d'ACO implémentée dans le

package ADE-4 (Thioulouse, *et al.*, 1997). Cette méthode permet d'extraire les composantes principales de la matrice (Gower, 1966). Cela permet de représenter nos *n* arbres dans un espace à deux dimension en croisant les facteurs les plus significatifs. On obtient une représentation où les arbres les plus proches topologiquement sont regroupés. Cette approche a l'avantage de ne pas considérer un arbre particulier comme référence.

2.3.2 ***Résultats***

2.3.2.1 Super-arbres basés sur 730 familles de gènes.

Après la sélection faite sur les arbres, 730 familles d'orthologues ont été identifiées. La distribution en taille de ces familles est montrée fig. 2.14. Seules les familles contenant plus de sept espèces ont été considérées. La forme caractéristique de la distribution illustre la nécessité d'une méthode permettant la prise en compte du signal phylogénétique apporté par les familles représentées dans moins de 45 espèces : les gènes ubiquitaires sont rares, et peu d'entre eux remplissent les critères que nous avons utilisés pour identifier les familles de bons orthologues. Par exemple, la fig. 2.15 montre la famille de gènes du facteur d'élongation EF-G (famille HBG000251) telle qu'elle apparaît dans HOBACGEN-CG. Seule la partie bactérienne de l'arbre est montrée. Bien que ce gène soit d'ordinaire considéré comme un bon marqueur phylogénétique (Bocchetta *et al.* 2000), l'on voit clairement que de nombreuses espèces possèdent deux ou trois paralogues anciens de ce gène, notamment *Synechocystis*. Cette famille n'a donc pas pu être retenue. Le même cas de figure est observé pour le paralogue ancestral de EF-G, le facteur d'élongation ET-Tu (Famille HBG016186). Ainsi, beaucoup de gènes ubiquitaires n'ont pu être considérés dans cette étude et ne devraient d'une manière générale être utilisés qu'avec une extrême prudence pour l'analyse phylogénétique.

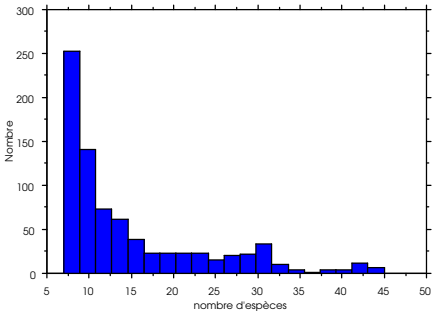


Fig. 2.14 : Distribution du nombre d'espèces contenues dans les 730 familles de gènes retenues pour l'analyse.

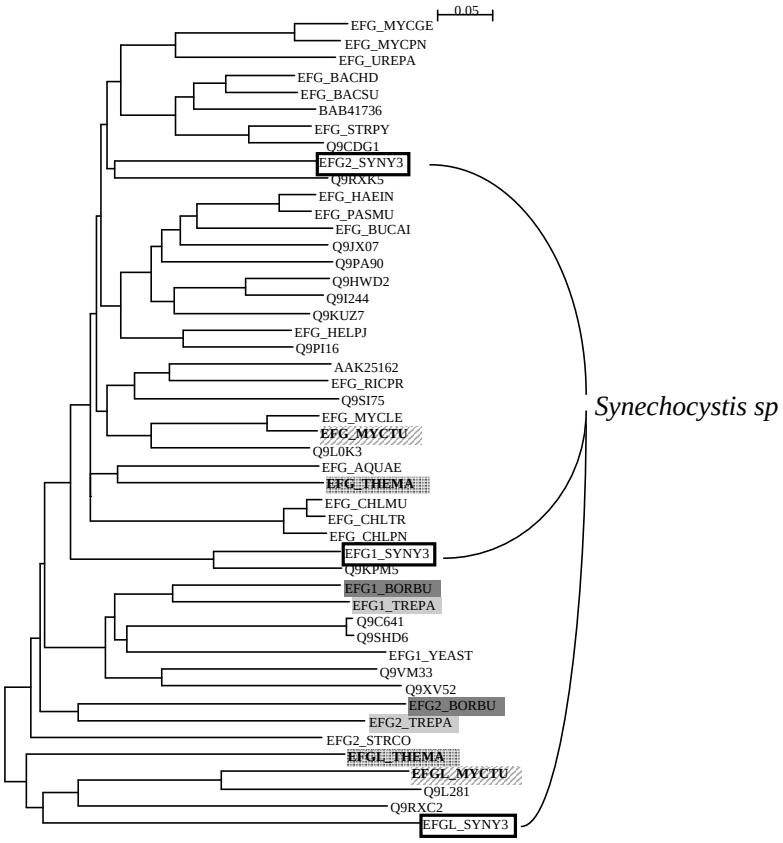


Fig. 2.15 : La phylogénie de la famille du facteur d'élongation EF-G telle qu'elle apparaît dans HOBACGEN-CG. Seule la partie bactérienne de l'arbre est montrée. La présence de plusieurs séquences très divergentes notamment de *Synechocystis* (SYNY3), *Borrelia* (BORBU), *Treponema* (TREPA), *Mycobacterium* (MYCTU) et *Thermotoga* (THEMA) suggère l'existence de paralogies cachées.

Les superarbres basés sur les 730 arbres de BIONJ + loi Gamma d'une part, et de Maximum de vraisemblance d'autre part sont présentés fig. 2.16 et fig 2.17 respectivement. Comme l'on pouvait s'y attendre, les trois domaines de la vie (archées, eucaryotes et bactéries) sont monophylétiques et bien soutenus. Les deux super-arbres sont remarquablement semblables, à l'exception des parties peu soutenues par les valeurs de bootstrap. Tous deux soutiennent la monophylie de la plupart des grands phylums archéens (euryarchaeotes et crénarchaeotes) et bactériens (gram-positives à bas G+C, gram-positives à haut G+C, spirochètes...) à l'exception notable des protéobactéries, dont le groupe des ε-protéobactéries change de position entre les deux arbres. Un groupement non trivial très bien soutenu est celui de *Deinococcus radiodurans* avec les bactéries gram-positives à haut G+C. Ainsi, il existe un signal fort pour la cohérence des grands groupes bactériens. Cependant, à l'instar de la plupart des études qui ont tenté d'utiliser les données génomiques afin de résoudre le problème de la phylogénie bactérienne, les relations entre ces grands groupes restent peu résolues. Cette difficulté à résoudre les branches profondes de l'arbre peut être interprétée comme étant due à la perte du signal phylogénétique sur les branches anciennes, ou, comme l'a suggéré Woese (Woese, 1987) à une radiation buissonnante entre les grandes

divisions bactériennes, mais également à l'augmentation de la probabilité d'avoir eu un transfert horizontal avec le temps de séparation. Également, la possibilité d'avoir des paralogies cachées dans les familles est probablement d'autant plus grande que les espèces sont distantes. Nous avons tenté de savoir si certains arbres particuliers étaient responsables de cet état de fait.

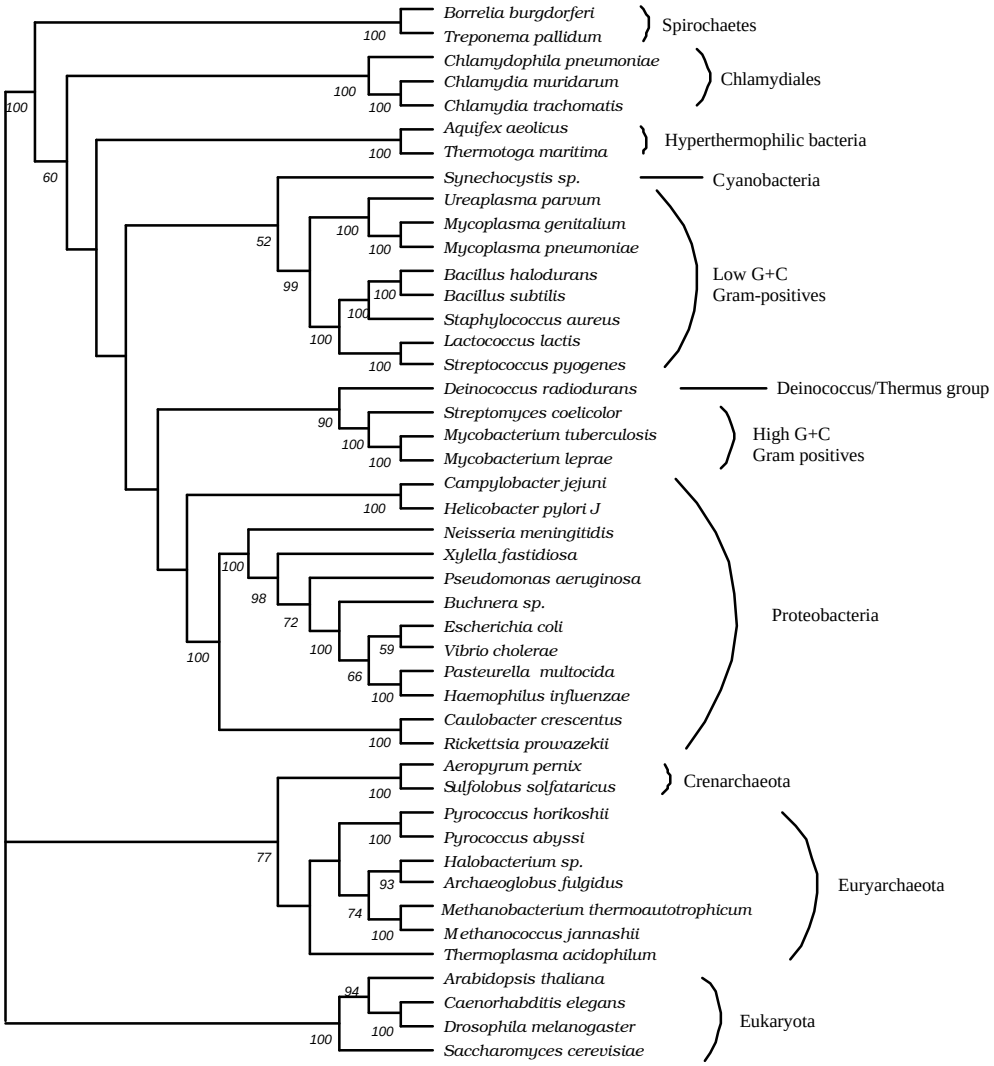


Fig. 2.16 : Superarbre basé sur 730 arbres construits avec la méthode BIONJ (Gascuel, 1997) en utilisant une distance basée sur une loi Gamma et un modèle de substitution JTT (Jones, *et al.*, 1992). Les valeurs de bootstrap (500 réplicats) supérieures à 50 % sont indiquées.

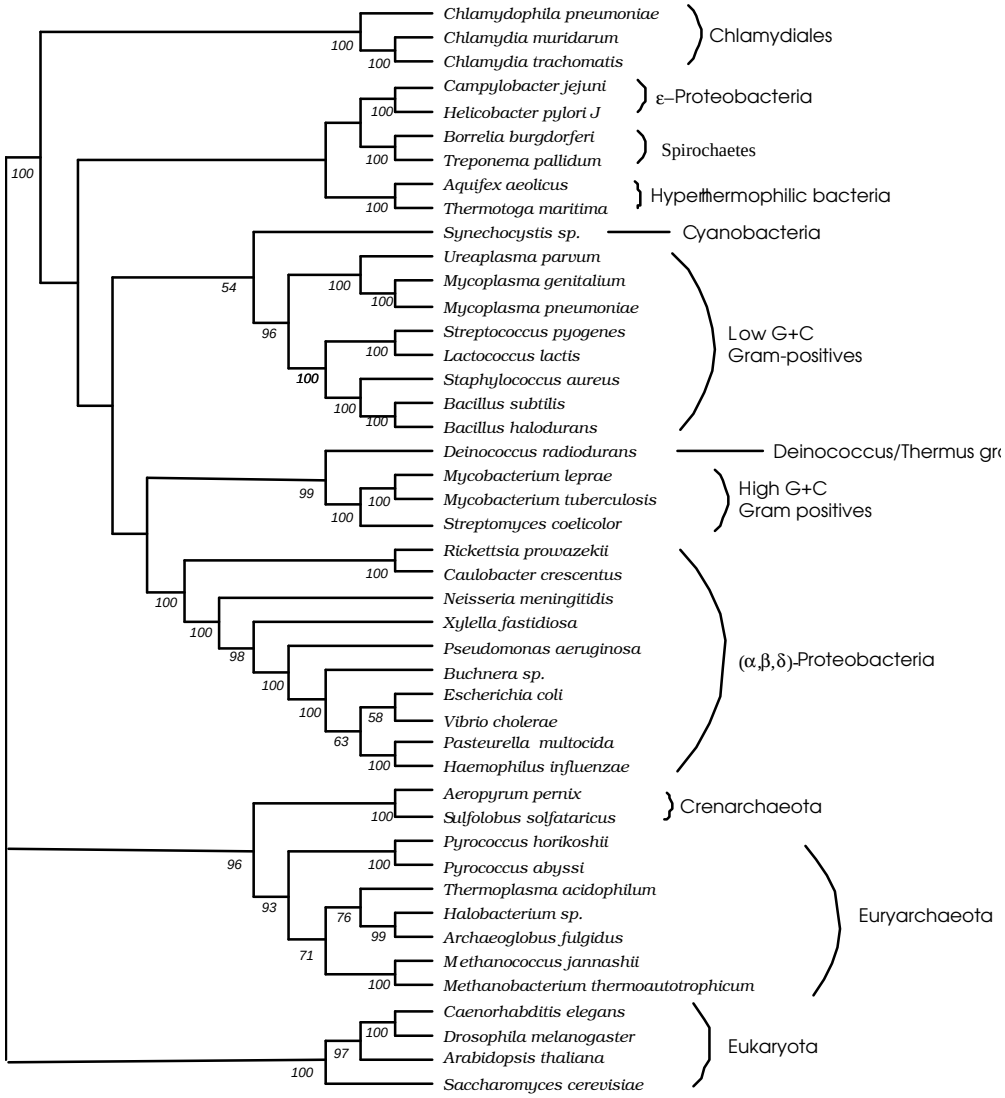


Fig. 2.17 : Superarbre basé sur 730 arbres construits avec la méthode de maximum de vraissemblance et un modèle de substitution JTT (Jones, *et al.*, 1992). Les valeurs de bootstrap (500 réplicats) supérieures à 50 % sont indiquées.

2.3.2.2 Comparaison des arbres de gènes.

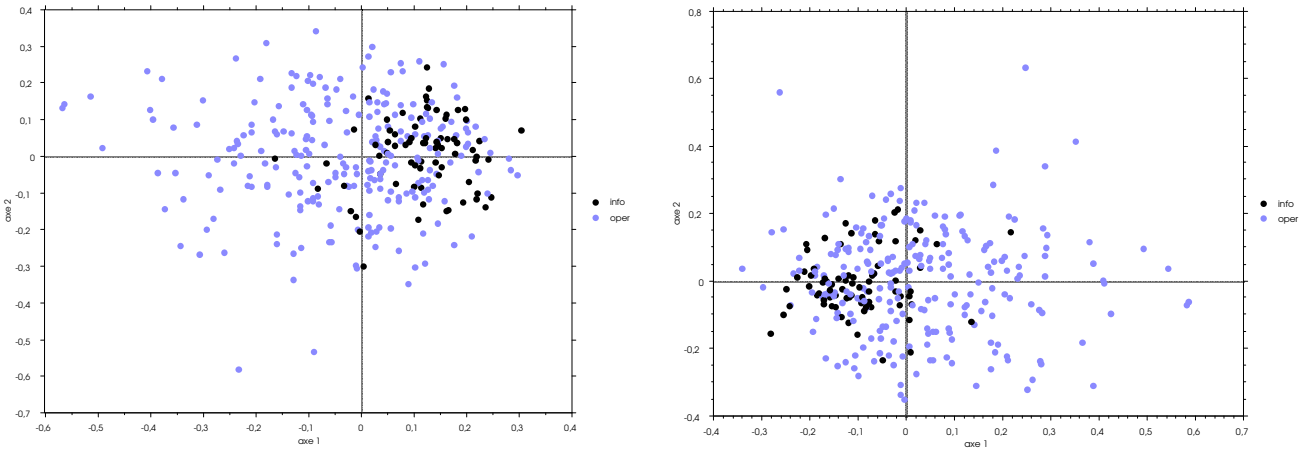


Fig. 2.18 : ACO sur les distances topologiques de 310 arbres contenant au moins 10 espèces bactériennes. Les ACO sont basées : à gauche, sur les arbres de BIONJ, et à droite, sur les arbres de Maximum de vraissemblance. Les points noirs représentent les gènes « informationnels » et les points gris, les gènes « opérationnels ». La majorité des gènes « informationnels » est concentrée dans la partie dense du nuage qui correspond à des arbres de topologies plus semblables entre elles.

L'hétérogénéité de taille des arbres fait que certains arbres ont des échantillonnages taxonomiques différents. Cela n'est pas gênant pour la reconstruction du superarbre tant qu'il existe des arbres pour « faire le pont ». Cependant, pour le calcul des distances topologiques entre arbres, ceci interdit de comparer l'ensemble des arbres car la matrice de distance peut contenir de nombreux « trous » que l'ACO ne peut gérer. Limiter le nombre de ces distances manquantes à moins de 10 % des cas permet de contourner ce problème en utilisant un estimateur des distances manquantes (voir section 2.3.1.7).

Pour limiter les distances manquantes à moins de 10 % des cas, nous avons choisi de ne comparer que les arbres contenant au moins dix espèces bactériennes. Trois cent dix arbres respectent ce critère et ont donc été comparés, puis la matrice a été analysée par ACO. Les résultats de cette analyse sont présentés fig. 2.18. Les deux premiers axes de l'analyse sont présentés. Dans les deux cas, le nuage de points comporte une région dense qui représente des arbres plus semblables entre eux. La bonne représentation dans cette région des gènes impliqués dans les fonctions liées à la gestion de l'information génétique est très remarquable et vient en confirmation des l'hypothèse de complexité émise par Jain *et al.* (Jain, *et al.*, 1999). Cependant, si les gènes informationnels tendent à contenir une information phylogénétique congruente, ils la partagent également avec de nombreux gènes dits

opérationnels. Ceci suggère que nous identifions ainsi des gènes impliqués dans de fonctions comparables aux fonctions informationnelles en terme de contraintes évolutives.

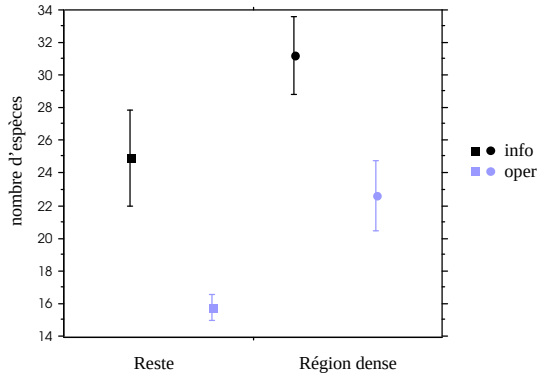


Fig. 2.19 : nombre moyen d'espèces dans les familles de la région dense et du reste du nuage de l'ACO pour l'analyse des arbres construits avec la méthode de BIONJ. Info : gènes informationnels ; oper : gènes operationnels. Les barres représentent 95 % d'intervalle de confiance.

Comme l'on pouvait s'y attendre, les gènes représentant ce « cœur de gènes » tendent à avoir une meilleure représentation phylogénétique: la fig. 2.19 montre que les gènes présents dans la région dense tendent à être présents dans plus d'espèces. Ainsi, le résultat de l'ACO peut s'interpréter de plusieurs manières. D'une part, l'on peut imaginer que les gènes les plus essentiels sont les mieux conservés à la fois du point de vue de leur présence dans une espèce et de leur séquence, ce qui correspond aux termes de l'hypothèse de complexité (Jain, *et al.*, 1999). Mais l'on peut également interpréter ce résultat comme un échec des méthodes de reconstruction à retrouver le bon arbre lorsque l'échantillonnage taxonomique est limité (Lecointre, *et al.*, 1993), ou bien encore comme une limite de notre méthode d'inférence des relations d'orthologie dans ces conditions. Dans ce cas, on ne peut tout à fait exclure que ce « cœur de gènes » ne soit non pas l'ensemble des gènes n'ayant pas ou peu subi de transferts, mais seulement ceux dont nous savons reconstruire l'histoire. Il n'en reste pas moins que ces gènes doivent pour cela subir des contraintes évolutives particulières qui suggèrent leur importance primordiale pour la cellule.

Les arbres regroupés dans ce nuage dense doivent supporter des topologies voisines, et de ce fait permettre de reconstruire un super-arbre des bactéries plus robuste. Nous avons donc repris ces arbres et reconstruit les super-arbres correspondants qui sont présentés fig. 2.20 et 2.21.

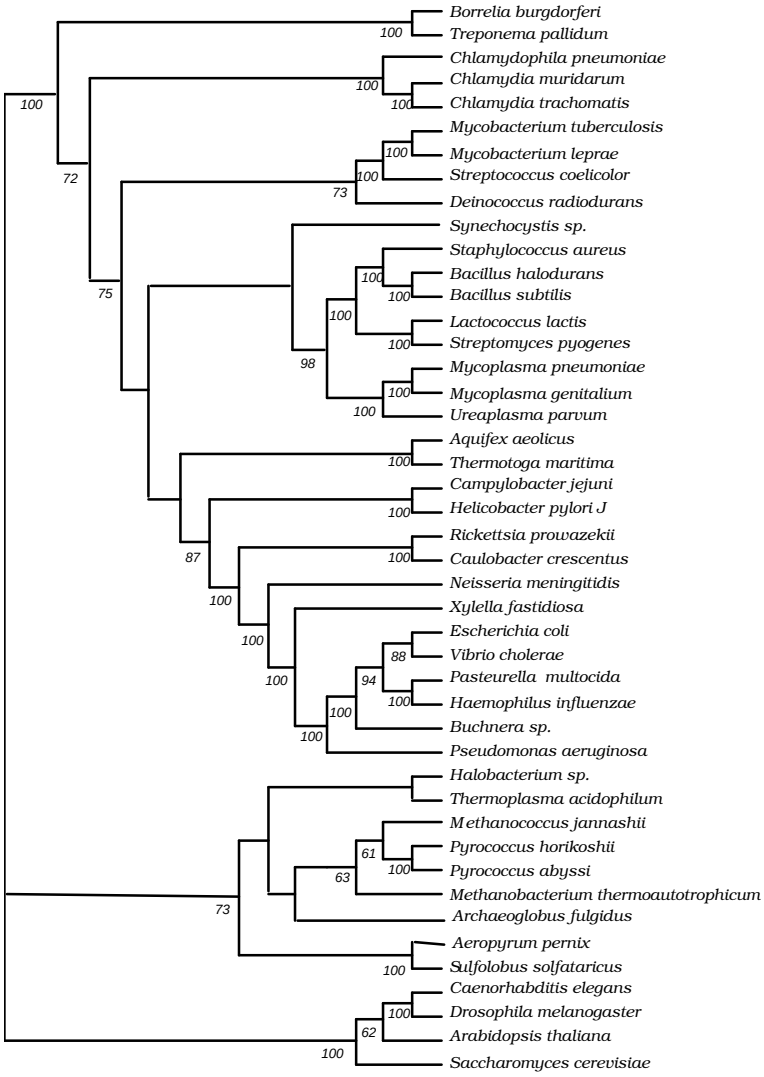


Fig. 2.20 : Superarbre basé sur 121 arbres construits avec la méthode BIONJ (Gascuel, 1997) en utilisant une distance basée sur une loi Gamma et un modèle de substitution JTT (Jones, *et al.*, 1992). Ces arbres ont été sélectionnés sur la base de leur appartenance à la région dense du nuage de l'ACO. Les valeurs de bootstrap (500 réplicats) supérieures à 50 % sont indiquées.

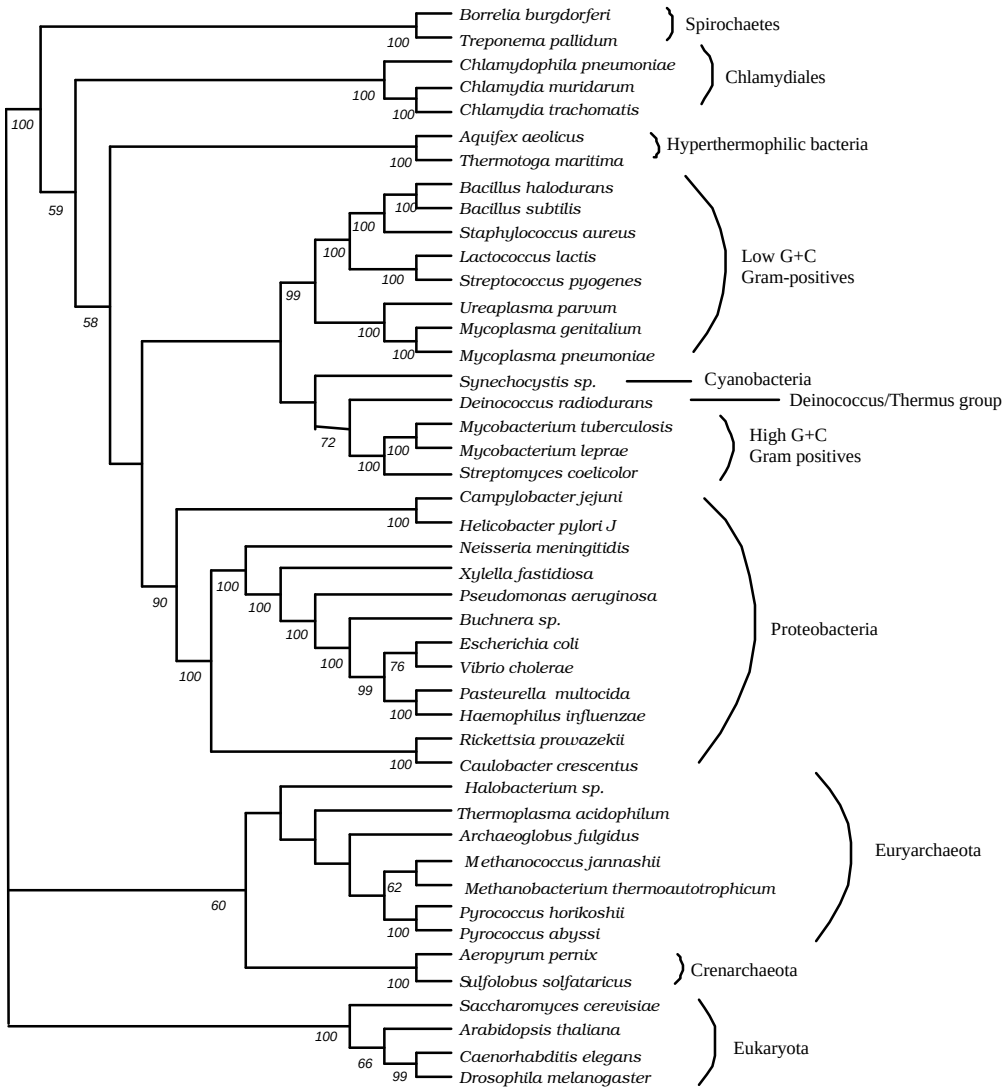


Fig. 2.21 : Superarbre basé sur 118 arbres construits avec la méthode de maximum de vraissemblance et un modèle de substitution JTT (Jones, *et al.*, 1992).. Ces arbres ont été sélectionnés sur la base de leur appartenance à la région dense du nuage de l'ACO. Les valeurs de bootstrap (500 réplicats) supérieures à 50 % sont indiquées.

Les topologies ainsi obtenues ne sont pas radicalement différentes des premiers super-arbres reconstruits en ce qui concerne les phylums bien connus. Cependant, un certain nombre de nœuds plus profonds sont ici bien soutenus, et notamment celui soutenant la monophylie des protéobactéries (ε-protéobactéries comprises). Le soutien statistique pour les bactéries occupant la position la plus basale a augmenté du fait de la sélection des arbres, et atteint même des valeurs supérieures à 70 %. Ainsi, les bactéries hyperthermophiles *Aquifex aeolicus* et *Thermotoga maritima* sont exclues de la position basale qu'elle occupent dans l'arbre de Woese (Woese, 1987). De même, la bactérie radio-résistante *Deinococcus radiodurans*, également considérée comme émergeant précocement dans l'arbre, est significativement groupée avec les bactéries gram-positives à haut G+C. Bien que surprenantes, ces positions vont dans le même sens que des travaux récents. Brochier *et al.* (Brochier, *et al.*, 2002), utilisant des protéines concaténées de bactéries ont montré la proximité de *Deinococcus* et des bactéries gram-positives. Bien que leur phylogénie ne soit pas racinée, ce qui interdit d'exclure la possibilité que *Deinococcus* se trouve à la base de l'arbre, ce résultat va dans le sens d'une relation encore ignorée entre ces deux groupes. Cette proximité a également été observée par Brown *et al.* (Brown, *et al.*, 2001). Malgré le fait que les bactéries gram-positives soient relativement proches dans l'un de nos deux arbres (celui basé sur les arbres ML), ceci implique que ce groupe n'est pas monophylétique et que la membrane externe aurait été perdue au moins deux fois indépendamment, d'une part par les bactéries gram-positives à bas G+C, et d'autre part par les haut G+C. Il est intéressant de noter que *Deinococcus* est positive à la coloration de Gram, mais possède une membrane externe. D'autre part, *Deinococcus* est proche de la bactérie thermophile *Thermus* (à tel point qu'un phylum les regroupant a été créé : le *Deinococcus/Thermus group*), qui est, elle, clairement Gram-négative.

La position non-basale des hyperthermophiles bactériens a déjà été abondamment discutée. Placées comme émergeant précocement et successivement dans l'arbre des bactéries de Woese (Woese, 1987) basé sur l'ARN ribosomal, *Thermotoga* et plus tard *Aquifex* ont été considérées comme la preuve du caractère hyperthermophile de l'ancêtre commun des bactéries. Cependant, cette position a été remise en doute notamment suite à la mise en évidence des contraintes fonctionnelles fortes s'exerçant sur l'ARN ribosomal des hyperthermophiles (Galtier, *et al.*, 1999) qui pourraient être à l'origine d'artefacts de reconstruction. De plus, une réévaluation récente de la phylogénie basée sur l'ARN ribosomal

soutient une position non basale des bactéries hyperthermophile (Brochier et Philippe, 2002). Ces résultats suggèrent donc une adaptation relativement récente de ces bactéries, notamment *via* l'acquisition par transferts horizontaux de certains gènes archéens nécessaires à la vie à haute température (Confalonieri, *et al.*, 1993; Forterre, 1995; Forterre, *et al.*, 2000).

La reconstruction du super-arbre bactérien en ne considérant que les séquences bactériennes pour la construction des arbres, avec la même sélection des familles (Fig. 2.22) donne un résultat identique à celui obtenu précédemment. Certains groupements, et notamment celui de *Deinococcus radiodurans* avec les Gram positives à haut G+C y sont mieux soutenus. Ce superarbre permet d'avoir une idée de la manière dont les groupes externes (archées et eucaryotes) influencent la topologie de la partie bactérienne de l'arbre, notamment *via* le phénomène d'attraction des longues branches. Le fait que la topologie soit peu ou pas affectée indique que l'effet de ces artefacts n'est pas suffisamment intense pour bouleverser le groupe interne. Cependant, le fait que cet arbre présente des soutiens statistiques supérieurs pour certains regroupements et notamment pour la proximité des bactéries gram-positives à haut G+C et de *Deinococcus*, ou pour la monophilie du groupe gram-positives bas G+C, cyanobactéries, *Deinococcus* et gram-positives haut G+C indique que certaines de ces espèces peuvent peuvent subir un phénomène d'attraction des longues branches dans les arbres contenant le groupe externe.

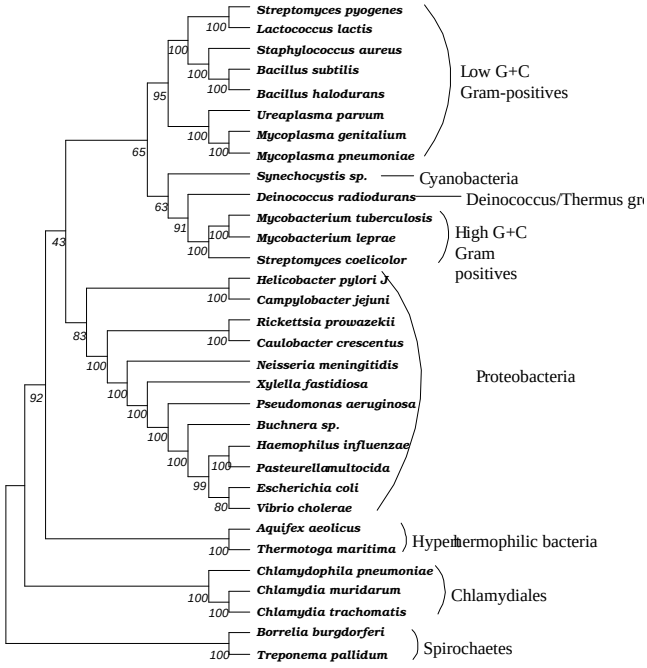


Fig. 2.22 : Superarbre basé sur les arbres de BIONJ reconstruits à partir des seules séquences bactériennes pour les 121 familles sélectionnées après l'ACO. La racine a été arbitrairement placée dans la même branche que celle suggérée par le superarbre raciné.

2.3.2.3 La partie archéenne de l'arbre

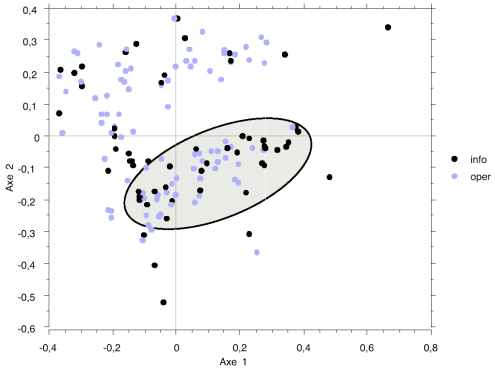


Fig. 2.23 : ACO sur les arbres construits à partir des séquences des seules archées par la méthode de BIONJ (JTT, loi gamma). L'existence d'une région dense est moins évidente qu'avec les bactéries. On peut cependant considérer que la région définie par l'ellipse est plus dense en gènes et contient une fraction importante des gènes informationnels

Le but premier de cette analyse n'était pas d'étudier la phylogénie des archées. De ce fait, la partie de l'arbre représentant le domaine archéen est peu résolue, notamment après la sélection des arbres par ACO, ce qui peut s'expliquer par le fait que toutes les familles majoritairement archéennes ont été exclues pour permettre les comparaisons d'arbres. Cependant, les superarbres basés sur les 730 familles sont relativement bien résolus pour cette partie de l'arbre. Nous avons reconstruit à partir de notre jeu de données des arbres ne contenant que les archées, en n'utilisant cette fois que la méthode BIONJ + loi Gamma. A partir des 149 arbres ainsi obtenus, une ACO a été effectuée (Fig. 2.23.). Bien que le résultat soit moins clair que pour les bactéries, une partie dense est là aussi identifiable et nous avons reconstruit l'arbre archéen correspondant aux 61 gènes ainsi sélectionnés. Les super-arbres avant et après sélection par ACO sont montrés fig. 2.24. Ils sont arbitrairement racinés entre les deux grands groupes d'archées (Creanarchaeotes et Euryarchaeotes), qui apparaissent nettement comme étant monophylétiques dans les super-arbres présentés plus haut. Ces arbres correspondent très largement à la phylogénie basée sur l'ARN ribosomal. Il sont en désaccord en ce qui concerne la position de *Thermoplasma*. Cependant, dans aucun des deux, la position de cette espèce n'est très fortement soutenue. La sélection après l'ACO permet cependant de mettre en évidence le clade regroupant *Halobacterium*, *Archaeoglobus* *Methanobacterium* et *Methanococcus*, mais il faut reconnaître que cette sélection se fait

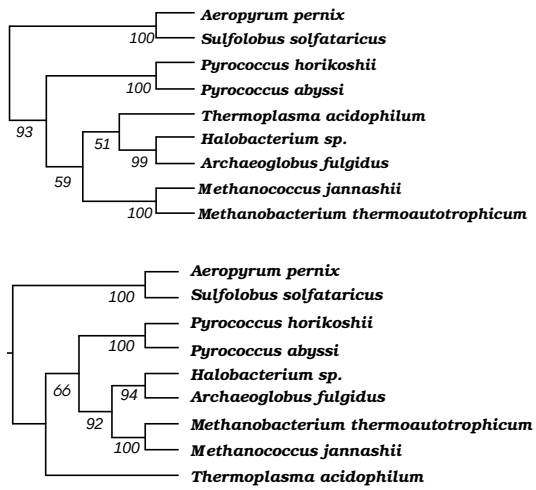


Fig. 2.24 : Superarbres des archées basés sur l'ensemble des 149 arbres (en haut) et sur les 61 arbres sélectionnés après l'ACO. La méthode utilisée ici est le BIONJ (JTT+loi gamma).

dans ce cas sur des critères plus subjectifs que dans l'étude sur les arbres bactériens, le nuage de points ne comportant pas de région dense bien marquée. L'un des problèmes majeurs de notre approche appliquée à l'arbre des archées est le nombre relativement faible d'espèces complètement séquencées, qui présage, comme nous l'avons noté plus tôt, qu'à la fois la reconstruction phylogénétique et les relations d'orthologie inférées sont incertaines.

Comme nous l'avons noté plus haut, aucune des familles contenant principalement des eucaryotes n'a été retenue pour cette analyse. Nous ne nous attarderons donc pas sur la partie eucaryote de l'arbre.

2.3.3 ***Discussion***

2.3.3.1 L'abondance des transferts horizontaux chez les bactéries

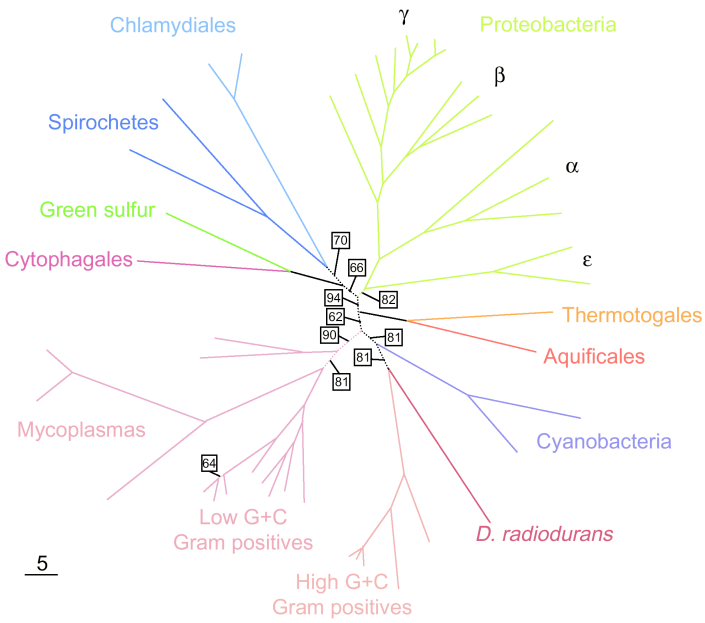
La question de savoir quelle est la quantité de gènes ayant subi des transferts horizontaux au cours de leur histoire est particulièrement compliquée à aborder. Notre étude s'attache à une catégorie très particulière de gènes : il s'agit des gènes suffisamment conservés et n'ayant subi que très peu de duplications au cours de leur histoire. Dans ce cadre très restreint, nous pouvons identifier près de 120 familles possédant une information suffisamment congruente sur la phylogénie bactérienne. Ces arbres représentent moins de la moitié des arbres testés pour leur ressemblance topologique. Cependant, il est difficile d'invoquer les transferts horizontaux plutôt que les artefacts de reconstruction ou les paralogies cachées pour expliquer cette majorité d'arbres incongruents. Nous avons noté plus haut que les arbres exclus de la partie dense du nuage tendaient à contenir moins d'espèces, ce qui suggère que des problèmes méthodologiques peuvent constituer une explication suffisante de ces incongruences, et qu'il faut faire attention à ne pas sur-interpréter ces résultats.

Il reste que les grands phylums bactériens sont facilement retrouvés par la méthode du superarbre, même sans faire de sélection sur la topologie des arbres. Ceci suggère que l'abondance des transferts entre ces groupes n'est pas suffisante pour dissoudre leur cohérence. La difficulté que nous avons à résoudre les noeuds profonds est plus probablement liée à la perte du signal phylogénétique entre ces groupes qu'à des transferts horizontaux.

La méthode telle que nous l'avons employée choisie de représenter chaque famille par un seul arbre, alors qu'il est possible que plusieurs arbres ne soient pas significativement différents du point de vue du critère phylogénétique choisi (*i.e.*, la longueur ou la vraisemblance de l'arbre). Le choix de ne coder que les nœuds soutenus par un *bootstrap* supérieur à 50 % réduit ce problème au niveau du codage de la matrice, mais pas au niveau de la comparaison des arbres. Malheureusement, il est difficile de concevoir un seul indice qui témoigne à la fois des différences topologiques entre arbre et du soutien de ces arbres.

2.3.3.2 Un consensus pour la phylogénie des bactéries ?

Un grand nombre d'études phylogénétiques basées sur les données génomiques échouent à donner des indications sur les relations de parenté existant entre les grandes divisions bactériennes (voir section 2.1). Cependant, l'analyse des résultats de deux méthodes indépendantes, celle de concaténation (Brochier, *et al.*, 2002) et de superarbre, permet d'identifier un certain nombre d'informations nouvelles sur la phylogénie des bactéries, les deux arbres étant remarquablement semblables (comparer par exemple les fig. 2.22 et 2.25). Notamment, la polyphylie des bactéries gram-positives est soutenue par les deux méthodes. Toutes deux proposent également qu'il existe un grand groupe monophylétique regroupant les cyanobactéries, *Deinococcus*, les gram-positives à haut G+C et les gram-positives à bas G+C (mycoplasmes compris) (sous réserve que la racine ne soit dans aucun de ces groupes). De même, la monophylie des protéobactéries est bien soutenue dans les deux méthodes. Ainsi, contrairement à ce que supposaient Teichmann et Mitchison (Teichmann et Mitchison, 1999), il semble que des



Fi. 2.25 : phylogénie non racinée des bactéries basée sur la concaténation de 57 familles de protéines impliquées dans la traduction des protéines. La topologie est très semblable à celle de la fig. 2.22 dont les divisions « green sulfur » et « cytophagales » sont absentes. Extrait de Brochier, *et al.*, 2002.

travaux utilisant l'information contenue dans les nombreuses familles de gènes disponibles aujourd'hui puissent apporter une nouvelle lumière sur les liens de parenté entre les grandes divisions et la phylogénie profonde des bactéries.

Dans leur ré-évaluation de la position de la racine dans la phylogénie des bactéries basée sur l'ARN ribosomal, Brochier et Philippe (Brochier et Philippe, 2002) ont montré que la racine pourrait se trouver dans la branche d'un groupe assez peu étudié de bactéries, celui des planctomycètes. Du fait du faible nombre de séquences disponibles pour ces organismes, ils sont malheureusement absents à la fois de notre étude et de l'étude de Brochier sur les protéines concaténées (Brochier, *et al.*, 2002). Cependant, il est remarquable que dans notre étude également, les bactéries hyperthermophiles n'aient pas non plus la position la plus basale. Une augmentation de la représentativité phylogénétique des données génomiques devrait à relativement cours terme permettre d'éclaircir ce point, notamment grâce à la méthode de superarbre.

2.4 Simulations sur le modèle du super-arbre

La méthode de super-arbre semble montrer une certaine robustesse aux transferts horizontaux notamment du fait de la similitude entre les super-arbres construits avant et après sélection sur des critères topologiques. Il est intéressant cependant d'évaluer dans quelle mesure les topologies aberrantes affectent la reconstruction phylogénétique par cette méthode. Pour ce faire, nous avons testé la capacité de la méthode de MRP à retrouver l'arbre vrai dans diverses conditions de perturbation des « arbres de gènes ».

2.4.1 ***Matériel et méthodes***

2.4.1.1 Perturbations à simuler

Les arbres de gènes peuvent ne pas représenter la phylogénie des espèces pour trois principales raisons : l'existence de paralogies cachées, les artefacts de reconstruction, et les transferts horizontaux. D'un point de vue topologique, ces trois types d'évènements ont les mêmes conséquences : le branchement d'une espèce où d'un groupe d'espèces à une position

erronée dans l'arbre. Nous avons donc décidé de simuler les perturbations directement au niveau topologique.

2.4.1.2 Simulation des arbres de gènes

Les simulations ont été conduites sur la base de deux arbres de références contenant chacun 32 espèces, afin d'étudier l'influence de la forme de l'arbre vrai sur la méthode de super-arbre. Le premier possède une topologie parfaitement symétrique, et le second une topologie asymétrique (Fig. 2.26.)

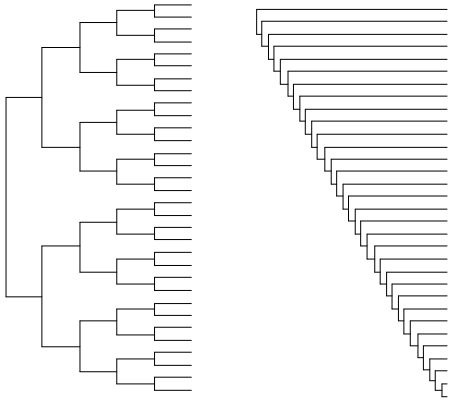


Fig. 2.26 : Les deux types d'arbres « vrais » dans les simulations. Chacun contient 32 espèces. A gauche, l'arbre est dit « symétrique » et à droite, « asymétrique »

Étant donné un arbre de référence, nous avons simulé soit 50, soit 100 « arbres de gènes » par les étapes suivantes :

-des pertes de gènes ont d'abord été simulées par suppression aléatoire de branches (internes ou externes). Les suppressions ont été faites de telle manière que pour un jeu d'arbres donné, la distribution des tailles soit approximativement normale avec une variance constante. Ainsi, nous ne considérerons comme variable que la moyenne de cette distribution.

- deux types de perturbations ont été séparément simulées, et sont décrites fig. 2.27 : le premier type de perturbation donne à chaque branche la même probabilité d'être choisie et déplacée à n'importe quel endroit dans l'arbre. Nous l'appellerons réarrangement global. Ceci correspondrait à des transferts horizontaux sans partenaires préférentiels. Par exemple, dans la fig. 2.27, le déplacement de la branche soutenant (AA, AB) correspond à la simulation de transferts de différentes espèces ancestrales à l'ancêtre commun à AA et AB. Dans ce cas, le

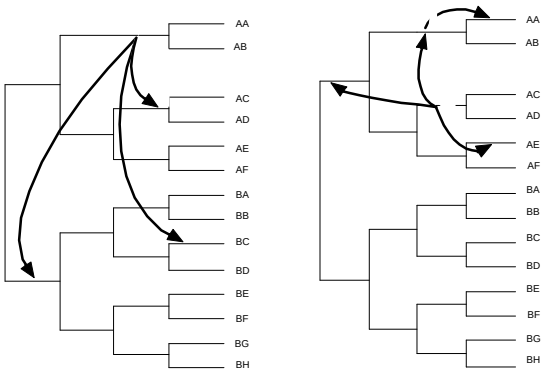


Fig. 2.27 : Les deux types de réarrangements simulés : réarrangements globaux (a gauche) et locaux (à droite).

paramètre retenu comme indice de la perturbation des arbres est le nombre de branches déplacées par espèce présente dans l'arbre.
Le second type de perturbations donne à chaque branche la même probabilité d'être choisie et d'être déplacée de sa position à une position voisine en traversant un nombre limité de nœuds. Nous l'appellerons réarrangement local. Ce type de réarrangement simule plus spécifiquement l'échec des méthodes de phylogénie à retrouver l'ordre de branchement correct des espèces à l'intérieur d'un groupe, mais également des transferts horizontaux impliquant préférentiellement des espèces voisines. Par exemple, dans la fig. 2.27, les perturbations représentées simulent l'incapacité de la méthode de reconstruction à retrouver la position correcte du groupe (AC, AD) au sein du groupe « A ». Ici, le paramètre retenu comme indice de la perturbation des arbres est le nombre de branches traversées par espèce présente dans l'arbre.

2.4.1.3 Comparaison entre arbres

Les paires d'arbres de gènes ont été comparées en utilisant l'indice de ressemblance topologique suivant (de même que dans le chapitre précédent, les deux arbres à comparer sont d'abord réduits aux taxons qu'ils ont en communs) :

$$I = \frac{b_c}{b_t}$$

Où b_c est le nombre de bipartitions communes aux deux arbres et b_t est le nombre total de bipartitions. Cet indice varie donc de 1 pour deux arbres identiques, à 0 pour deux arbres ne possédant aucune bipartition commune.

2.4.1.4 Calcul des super-arbres

Les super-arbres sont reconstruits à partir de 50 ou 100 « arbres de gènes ». Pour chaque valeur du nombre moyen d'espèces par arbre (noté sp dans les figures), 500 super-arbres sont reconstruits sur une gamme continue du paramètre de perturbation (noté tr dans les figures).

2.4.2 ***Résultats et discussion***

Les résultats des simulations sont présentés dans les figures 2.28 et 2.29.

Pour faciliter la représentation des résultats, les paramètres ont été regroupés en classes. Ainsi, en abscisse des graphes, les résultats sont regroupés par intervalles de la moyenne de la taille des arbres simulés. De même, les moyennes des indices de perturbation ont été regroupés en classes.

D'une manière générale on observe, comme on pouvait s'y attendre, d'une part que les arbres contenant beaucoup d'espèces permettent plus facilement de retrouver la référence à 32 espèces ; d'autre part que plus on utilise d'arbres, plus on a de chance de retrouver la topologie de référence (les similarités pour les super-arbres basés sur 100 arbres sont toujours supérieures à celles basées sur 50 arbres) ; et enfin que plus les arbres sont perturbés, plus il est difficile de retrouver la référence. Ces trois effets sont cumulatifs. Cependant, l'on voit que les effets de la taille des arbres et des perturbations peuvent être compensés par la prise en compte d'un très grand nombre d'arbres.

Il est intéressant de noter qu'il existe une forte sensibilité de la méthode à la forme de l'arbre de référence : les arbres asymétriques sont beaucoup plus difficiles à retrouver.

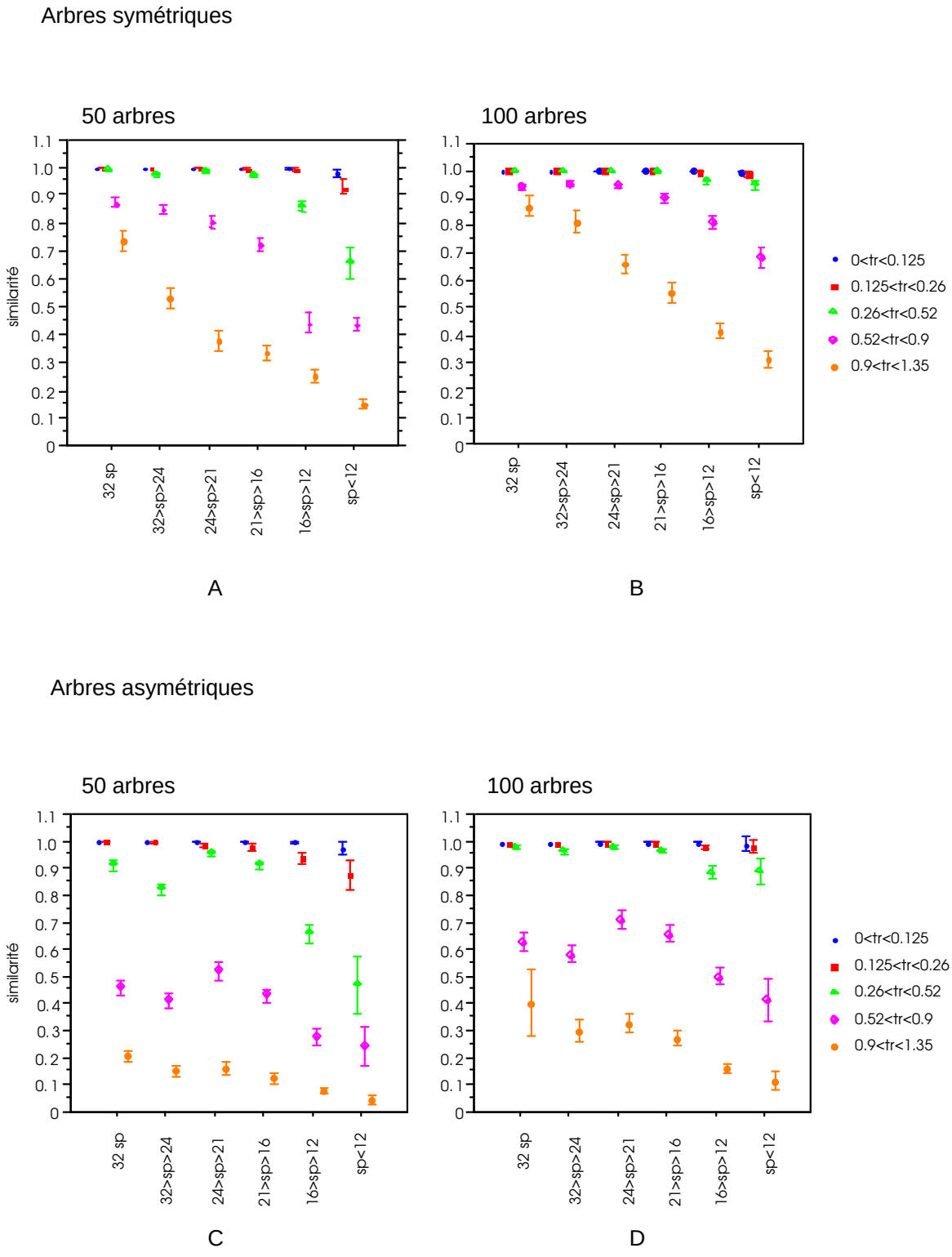


Fig. 2.28 : Résultats des réarrangements globaux. Les graphes présentent la moyenne des similitudes topologique (indice *I*) entre le superarbre reconstruit et l'arbre vrai pour différents paramètres de perturbation (moyenne(transfert/sp/arbre) noté tr) et de nombres moyens d'espèces dans les « arbres de gènes » (sp). Les résultats sont présentés pour des tests utilisant 50 (A et C) ou 100 (B et D) « arbres de gènes » simulés sur la base d'« arbres vrais » symétriques (A et B) ou asymétriques (C et D).

2.4.2.1 Réarrangements globaux

La méthode de super-arbre montre une robustesse intéressante au taux de transferts. Par exemple, en utilisant 50 arbres contenant en moyenne 16 à 21 espèces ($16 < sp < 21$), même avec des perturbations supérieures à un transfert pour quatre espèces ($0,26 < tr < 0,52$), plus de 95 % des bipartitions de l'arbre de référence sont systématiquement retrouvées. Si l'on utilise 100 arbres, ce chiffre est proche de 100 %. Dans la classe de taille d'arbre inférieure ($12 < sp < 16$), l'on voit qu'au même niveau de perturbation, le passage de 50 à 100 arbres permet de faire passer le pourcentage de bipartitions correctes de moins de 90 % à plus de 95 %.

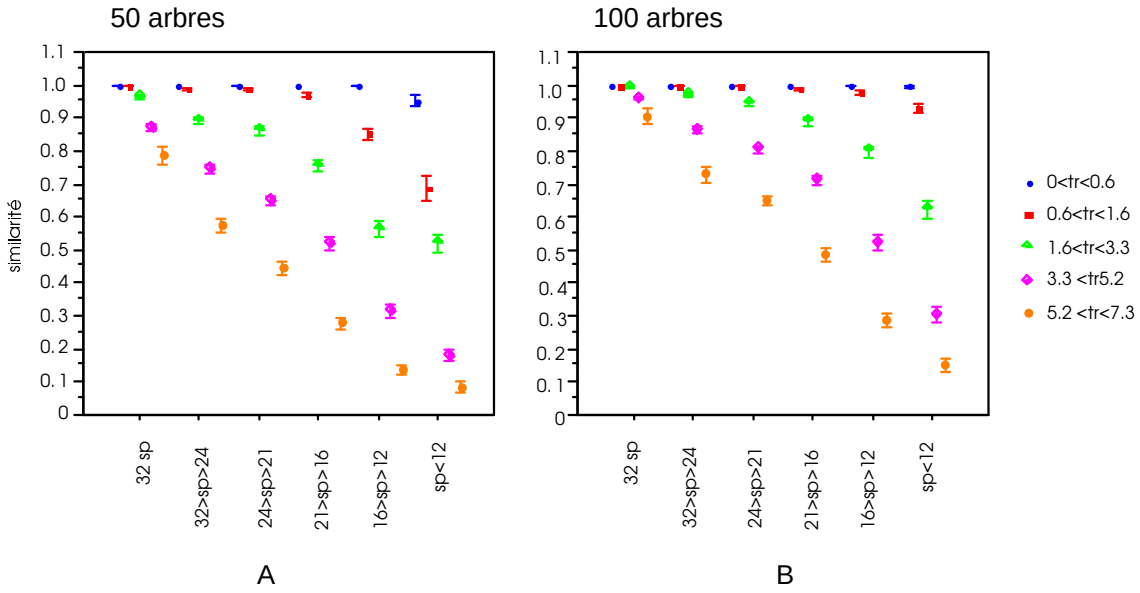
Les différences de résultats de la méthode en fonction de la topologie de l'arbre de référence ne sont véritablement visibles que lorsque les taux de transferts sont élevés. Pour les arbres asymétriques, la méthode MRP donne des résultats très comparables lorsque les taux de perturbation sont très forts ($tr > 0,52$) quels que soient le nombre et la taille des arbres. Dans de tels cas, il semble complètement vain de tenter de reconstruire la topologie de référence.

2.4.2.2 Réarrangements locaux

De même que pour les réarrangements globaux, la méthode est assez résistante aux réarrangements locaux. Il faut simuler plus d'un transfert pour deux espèces pour que, avec des tailles d'arbres raisonnables ($12 < sp < 16$ par exemple), la méthode ne retrouve pas l'arbre de référence à chaque fois. Il est particulièrement notable que, même avec des taux de réarrangement supérieurs à 1, la méthode peut retrouver plus de 90 % des bipartitions si les arbres sont suffisamment grands et nombreux (100 arbres, $21 < sp < 24$).

La sensibilité au nombre d'espèces dans les arbres utilisés est ici plus marquée que dans les réarrangements globaux. Ceci peut probablement s'expliquer par le fait que pour des arbres dans lesquels beaucoup d'espèces ont été supprimées, même les réarrangements locaux correspondent à des différences profondes du point de vue phylogénétique.

Arbres symétriques



Arbres asymétriques

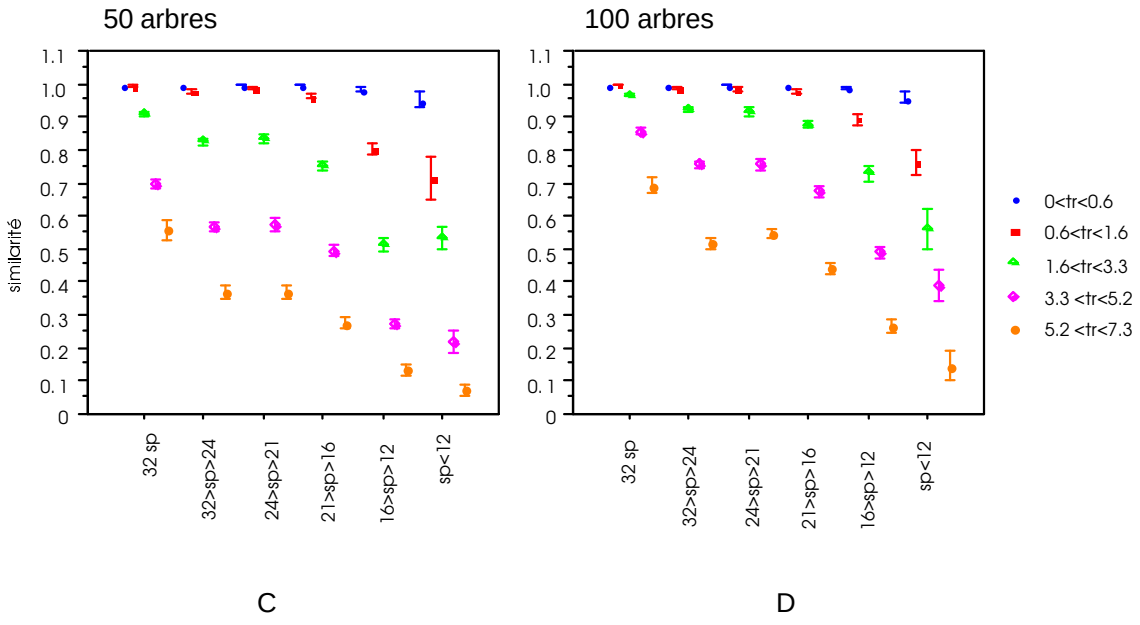


Fig. 2.29 : Résultats des réarrangements locaux. Les graphes présentent la moyenne des similitudes topologique (indice *I*) entre le superarbre reconstruit et l'arbre vrai pour différents paramètres de perturbation (moyenne(transfert/sp/arbre) noté tr) et de nombres moyens d'espèces dans les « arbres de gènes » (sp). Les résultats sont présentés pour des tests utilisant 50 (A et C) ou 100 (B et D) « arbres de gènes » simulés sur la base d'« arbres vrais » symétriques (A et B) ou asymétriques (C et D).

2.4.2.3 Relation avec la similitude des arbres de gènes

La fig. 2.30 montre la décroissance de la similitude moyenne entre les « arbres de gènes » en fonction du taux de perturbation des arbres. La décroissance observée dépend à la fois de la forme de l'arbre, du nombre d'espèces présentes dans les arbres et du mode de perturbation. Par exemple, l'indice moyen de similitude topologique semble être toujours au moins légèrement supérieur dans les arbres à 18 espèces, en comparaison des arbres à 32 espèces. Ceci est particulièrement marquant pour les arbres asymétriques subissant des réarrangements globaux : très peu de réarrangements (de l'ordre de 0,05 transfert/sp/arbre) suffisent à faire chuter la similitude moyenne des arbres contenant 32 espèces à moins de 0,3, cependant que le même taux de réarrangement ne fait tomber qu'à 0,6 la similitude moyenne des arbres à 18 espèces. Ceci peut s'expliquer par le fait que les arbres à 18 espèces peuvent montrer des échantillonnages taxonomiques différents et lorsque les espèces ayant subi un transfert ne sont pas communes aux deux arbres comparés, le réarrangement ne diminue pas la similitude entre les arbres. Ainsi, du fait de cette dépendance au nombre d'espèces dans les arbres, on peut s'attendre à ce qu'il n'existe pas de relation simple entre la similitude observée entre les arbres et la probabilité de reconstruire le bon arbre par la méthode. Cependant, les grosses différences entre les courbes de similitude correspondent, au moins pour les réarrangements globaux, à des taux de transferts où la méthode MRP retrouve presque toujours le bon arbre (voir fig. 2.28). Il est particulièrement remarquable de constater que

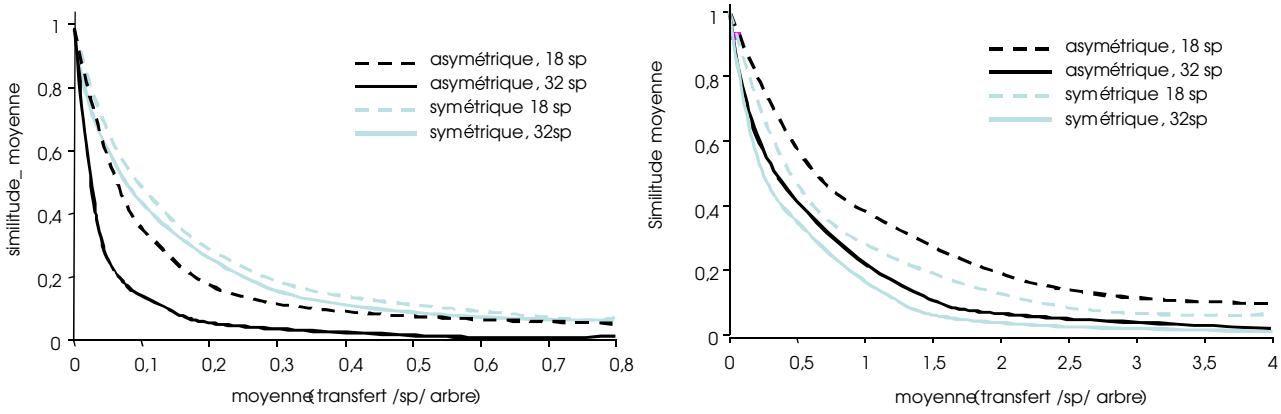


Fig. 2.30 : Décroissance observée de l'indice de similitude topologique entre arbres avec le degré de perturbation des arbres (moyenne(transfert/sp/arbre) noté tr dans les figures précédentes) pour le réarrangements globaux (gauche) et locaux (droite) et pour différents types d'arbres.

pour les réarrangements globaux, on peut reconstruire le bon arbre avec une grande efficacité même avec des arbres ayant une similitude moyenne inférieure à 0,2.

Le cas des réarrangements locaux est assez semblable de ce point de vue. Cependant, la probabilité de reconstruire le bon arbre décroit plus rapidement en fonction de la similitude entre les « arbres de gènes » : on observe une faible tendance à ne pas retrouver le bon arbre, avec des « arbres de gènes » ayant pourtant des similitudes entre eux de l'ordre de 0,3-0,4 et ce, en particulier lorsque l'arbre de référence est asymétrique (fig. 2.30). Ceci est particulièrement intéressant car ce mode de réarrangement est probablement plus réaliste que les réarrangements globaux. Par exemple, pour revenir à la problématique de l'arbre des procaryotes, il est probable que les transferts se font plus fréquemment dans la nature entre bactéries ou entre archées plutôt qu'entre ces deux domaines. De ce fait, lorsque l'on compare les arbres, certaines bipartitions qui correspondent par exemple à la bipartition archées/bactéries sont retrouvées dans presque tous les arbres. Il peut donc exister entre les arbres une « similitude basale », quasiment indépendante du taux de transfert, uniquement due au fait que les transferts se font localement dans l'arbre. On peut donc envisager que même avec des similitudes moyennes relativement fortes entre « arbres de gènes », la méthode MRP échoue à retrouver certaines branches.

2.4.2.4 Réalisme des simulations

Comme toutes expériences de simulation, celles décrites ici souffrent d'un manque de réalisme. Nous avons d'abord considéré que les transferts horizontaux pouvaient concerner toutes les espèces avec la même probabilité. Il semble au contraire que, dans la nature, certaines espèces soient plus aptes que d'autres à intégrer de l'ADN étranger dans leur génome. De même, dans les cas des réarrangements locaux, censés simuler également un certain nombre d'artefacts de reconstruction phylogénétique, on connaît très bien des cas d'espèces dont une grande partie des gènes subissent des taux d'évolution important, ce qui rend leur branchement dans les arbres très aléatoires. Enfin, nous avons largement considéré que les transferts ne se faisaient pas entre partenaires préférentiels. Les réarrangements locaux modélisent un cas particulier de ce phénomène (le cas d'échanges préférentiels entre espèces proches), mais ne permettent pas d'étudier les cas où les échanges se font plus fréquemment entre des espèces éloignées dans l'arbre mais partageant un même habitat.

Cependant, de même que certains résultats de ces simulations peuvent paraître triviaux (« il vaut mieux, pour reconstruire un super-arbre, disposer de beaucoup d'arbres peu perturbés et contenant beaucoup d'espèces que l'inverse »), l'on peut d'avance prédire que l'impact de ces phénomènes est forcément important. Mais le plus problématique est certainement l'existence de transferts impliquant préférentiellement certaines espèces. De ce fait, il est important de tenter de limiter ce phénomène dans les données. Par exemple, les transferts systématiques entre espèces peuvent parfois être identifiés comme dans le cas des eucaryotes et des α-protéobactéries, ou encore des bactéries et archées hyperthermophiles. Il apparaît donc souhaitable, comme nous l'avons fait précédemment, d'enlever les familles correspondantes ou du moins les gènes concernés.

En ce qui concerne les espèces subissant des taux d'évolution forts, il est possible que la méthode puisse s'en accommoder, dans la mesure où certaines précautions sont prises au niveau des arbres de gènes. Il est évident que si la plupart des arbres contiennent un regroupement de deux espèces en réalité non apparentées, le super-arbre aura tendance à les placer ensemble. Mais les artefacts de reconstruction sont spécifiques des gènes, et de l'échantillonnage considéré. Or, dans une approche de super-arbre, où par définition les échantillonnages taxonomiques peuvent être variables d'un arbre à l'autre, il est probable que les regroupements artefactuels systématiques d'espèces soient relativement rares et en tout cas minoritaires face aux regroupements légitimes (une espèce ayant un taux d'évolution fort aura tendance à se grouper avec une autre espèce évoluant rapidement, mais qui pourra être différente d'un arbre à l'autre). Cette spéculation est difficile à tester par des simulations. Cependant, l'application aux données est plutôt rassurante de ce point de vue. Par exemple, les espèces du groupe des mycoplasmes sont connues pour avoir des taux d'évolution très forts (Woese, *et al.*, 1984), qui les placent très souvent à la base des bactéries aussi bien dans les arbres basés sur des gènes uniques (Gupta, 1998b; Klenk, *et al.*, 1999) qu'avec des approches multi-genes (Teichmann et Mitchison, 1999; Hansmann et Martin, 2000; Lin et Gerstein, 2000). Cependant, elles se positionnent dans le super-arbre avec les autres bactéries gram-positives à bas G+C, ce qui semble être leur vraie place (Woese, *et al.*, 1984).

2.4.2.5 Avantages et inconvénients de la méthode de super-arbre

Ainsi, il semble que les propriétés de la méthode MRP soient plutôt bonnes, dans la mesure où elle est capable de retrouver très efficacement le bon arbre à partir d'un jeu d'arbres perturbés de manière non biaisée. Le codage en une matrice de parcimonie permet de concentrer toute l'information phylogénétique en un nombre restreint de caractères informatifs. De ce fait, elle présente de grands avantages par rapport aux autres méthodes de super-arbres basées sur la combinaison d'arbres à quatre espèces (quartets), puisqu'elle permet d'obtenir un résultat rapidement. En effet, la décomposition en quartets de centaines d'arbres puis leur combinaison en un super-arbre pose des problèmes algorithmiques complexes qui ne sont solubles qu'avec des temps de calcul importants (Bryant et Steel, 2001).

Comme nous l'avons noté plus tôt, la méthode de MRP ne permet pas de prendre en compte le fait que plusieurs topologies puissent être quasiment équiprobables pour un alignement. Si le seuil des valeurs de bootstrap évite de prendre en compte des regroupements non significatifs, il ignore par la même occasion une information que les méthodes de concaténation pourraient, elles, mettre à profit : si certains nœuds ne sont résolus (n'ont pas un support supérieur à 50 %) dans aucun des arbres, le super-arbre ne contiendra aucune information sur ces nœuds alors que la méthode de concaténation pourrait théoriquement le faire. C'est typiquement ce qui se produit lorsque, par exemple, l'on reconstruit le super-arbre correspondant au jeu de 23 gènes de Brown, *et al.*, 2001 (résultats non présentés) : si la topologie obtenue par MRP est très semblable à celle obtenue par la méthode de concaténation, elle présente, contrairement à la fig. 2.1 un soutien très faible pour les nœuds profonds.

La question de savoir ce qui, de la méthode de concaténation des séquences ou des méthodes de super-arbre, est le mieux adapté pour la reconstruction phylogénétique basée sur les séquences reste cependant complexe. D'une part, on considère souvent que le signal phylogénétique de deux alignements bruités peut émerger après leur concaténation, et il est vrai que les arbres basés sur des concaténats de gènes présentent souvent une bonne résolution. Cependant, il semble que l'on puisse par cette méthode trouver des arbres faux (du moins du point de vue des auteurs) très bien soutenus (cf. notamment Brown, *et al.*, 2001). D'autre part, même dans une approche plus rigoureuse où les alignements sont choisis en

fonction des topologies qu'ils soutiennent (Brochier, *et al.*, 2002; Matte-Tailliez, *et al.*, 2002), le concaténat de séquences soutient un arbre très différent de ceux qui ont déterminé le choix des gènes à concaténer. En effet, dans les ACP, les points représentant les concaténats des familles identifiées comme n'ayant pas subi de transferts sont souvent éloignés du nuage de point de ces mêmes familles prises individuellement. Ceci suggère que les méthodes de reconstruction phylogénétique disponibles actuellement sont incapables de prendre en compte la diversité des modalités d'évolution des différents gènes. Dans le même esprit, Bapteste *et al.* (Bapteste, *et al.*, 2002) ont montré qu'il était préférable, dans une approche de maximum de vraisemblance, de considérer les gènes un à un plutôt que concaténés et de maximiser la somme des vraisemblances des alignements plutôt que la vraisemblance de l'alignement concaténé. Ainsi, dans l'approche par ACP, il serait intéressant non pas d'étudier le concaténat des familles n'ayant pas subi de transfert, mais d'identifier les topologies responsables de ces regroupements en traçant le cercle des corrélations.

2.5 Tentative d'amélioration des critères de sélection des gènes à concaténer

Le point de savoir comment choisir les données à prendre en compte pour la reconstruction phylogénétique a été longuement discuté dans le débat opposant les partisans de la prise en compte simultanée de toutes les données disponibles, morphologiques et moléculaires, pour résoudre un problème phylogénétique (« total evidence ») et ceux d'un traitement indépendant des données. Si la combinaison des données a souvent montré une bonne capacité de résolution des arbres, cette approche suppose que les méthodes phylogénétiques sont consistantes, c'est-à-dire qu'elles tendent vers l'arbre vrai quand la quantité de données augmente. Cependant, plusieurs travaux suggèrent que les méthodes de reconstruction ne sont pas consistantes lorsque les données combinées sont hétérogènes et insistent sur la nécessité de ne combiner que des données congruentes (Cunningham, 1997b; Cunningham, 1997a).

2.5.1 ***Le test ILD (« Incongruence Length Difference »)***

Outre le critère proposé par Brochier *et al.* 2001 pour déterminer quels alignements soutiennent des arbres semblables, d'autres tests de congruence des données ont été proposés, dont le plus représentatif est certainement le test d'ILD (pour « Incongruence Length Difference ») lié à la méthode de parcimonie (Farris, *et al.*, 1994). Ce test a d'abord été proposé pour quantifier les conflits qui peuvent exister entre des données de sources différentes comme des données de séquences de différents compartiments cellulaires (noyau ou organites), des données de polymorphismes (RFLP, RAPD...), ou encore des traits morphologiques. Appliqué aux alignements de séquences, ce test pourrait constituer un outil intéressant pour détecter les alignements portant des informations aberrantes par rapport aux autres.

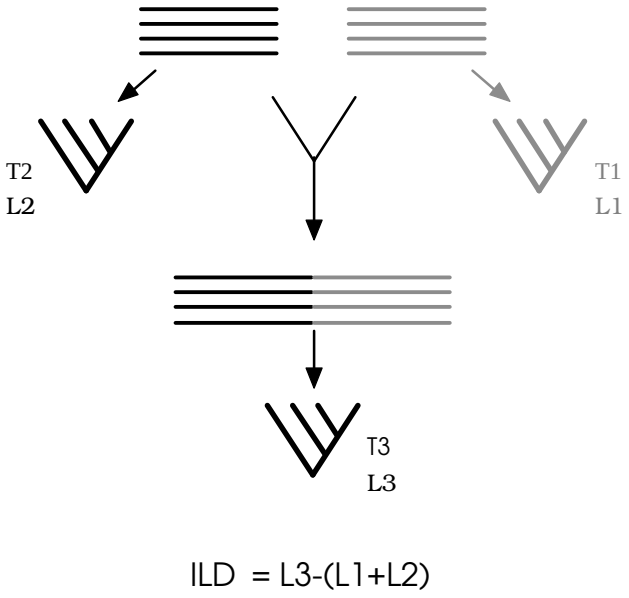


Fig. 2.31 : Le principe du test ILD. Voir détails dans le texte.

Nous allons détailler rapidement le principe de cette méthode : soient deux matrices de caractères. Chacune de ces matrices peut donner un ou plusieurs arbres optimisant le critère de parcimonie, c'est-à-dire nécessitant d'inférer le minimum d'évènements évolutifs. Ainsi, à la matrice 1 correspond un ensemble d'arbres T1 qui ont tous la longueur minimale `L1`. De même pour la matrice 2. Si ces deux matrices soutiennent au moins une topologie en commun, le concaténat des deux soutient un ensemble d'arbres T3 qui correspond à l'intersection de T1 et T2, et dont la longueur `L3` est égale à la somme de `L1` et `L2` (Fig. 2.31). Farris *et al.* (1995) définissent l'indice d'ILD comme la différence entre la longueur de l'arbre construit à partir de la matrice concaténée (`L3`) et la somme des longueurs des arbres construit à partir des matrices initiales (`L1` et `L2`). L'ILD est nulle si les matrices sont congruentes, et positive sinon. Cet indice peut alors être interprété comme le nombre d'évènements évolutifs supplémentaires que l'incompatibilité des deux matrices nécessite d'inférer. L'indice ILD correspond donc à un nombre

d'évènements évolutifs inférés, et son importance est relative à la taille des matrices. Farris *et al.* (Farris, *et al.*, 1994) proposent donc un test statistique qui consiste à répartir aléatoirement les sites présents dans les deux matrices en deux nouvelles matrices de même taille. Cette étape, renouvelée un certain nombre de fois, permet d'obtenir une distribution des valeurs d'ILD sous l'hypothèse de congruence des deux matrices. Le positionnement de la valeur d'ILD réelle dans cette distribution permet de savoir si les matrices sont significativement incongruentes.

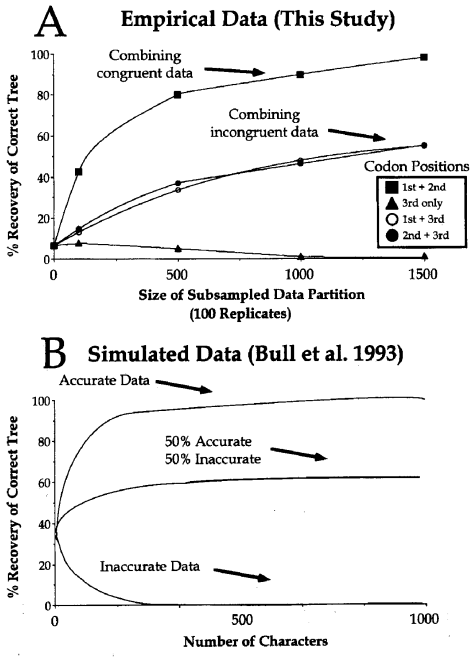


Fig. 2.32 : l'effet de la combinaison de données sur les reconstructions phylogénétiques, ou faut-il préférer la qualité à la quantité des données. A la fois les tests empiriques (A) et les simulations (B) montrent que la prise en compte de toutes les données disponibles n'est pas une solution et qu'il faut rechercher les données congruentes. Extrait de Cunningham, 1997b

Ce test a été appliqué à de nombreux problèmes phylogénétiques (Lecointre, *et al.*, 1998; Dolphin, *et al.*, 2000; Yoder, *et al.*, 2001; Brown, *et al.*, 2002) et plusieurs auteurs on tenté d'évaluer sa sensibilité à différents facteurs, à la fois sur des données réelles et simulées. Il a ainsi été montré que la prise en compte de la congruence des données à concaténer pouvait fortement améliorer les résultats (Cunningham, 1997b; Cunningham, 1997a) et que le test ILD constituait l'une des meilleures méthodes pour cela (Cunningham, 1997a). Cependant, plusieurs défauts de ce test ont été pointés, comme sa sensibilité à la différence de taille des matrices comparées et notamment sa tendance à surestimer la congruence lorsqu'une des deux matrices est très supérieure à l'autre en taille (Dowton et Austin, 2002). D'autre part, l'on sait également que la congruence des données n'est pas nécessairement synonyme d'augmentation du pouvoir de résolution car des jeux de données peu résolutifs peuvent être congruents entre eux (Cunningham, 1997b). Récemment, Darlu et Lecointre (Darlu et Lecointre, 2002) ont testé la sensibilité du test à différents paramètres d'évolution des séquences. Ils montrent notamment que le test ILD est sensible à l'hétérogénéité des taux d'évolution entre sites et au nombre de sites dans les matrices. Particulièrement, lorsque les matrices contiennent peu de caractères et que les taux d'évolution sont hétérogènes, le test a tendance à trouver les matrices congruentes même lorsqu'elles ne le sont pas.

Comme le note Cunningham (Cunningham, 1997b), les propriétés sur lesquelles repose le test ILD ne sont pas propres à la parcimonie et l'on peut en théorie l'étendre aux autres méthodes de reconstruction phylogénétique. Notamment, les méthodes de distance utilisant le critère d'évolution minimum, comme le Neighbor-Joining (NJ) ou BIONJ (Gascuel, 1997) minimisent, de même que la parcimonie, le critère « taille de l'arbre ». Nous avons donc expérimenté, en collaboration avec Marina Zelwer, l'extension de ce test à l'algorithme de BIONJ (Gascuel, 1997).

2.5.2 ***Adaptation de l'ILD aux méthodes de distance***

Le test ILD est basé sur plusieurs propriétés qu'il convient d'essayer de conserver pour l'extension aux méthodes de distance.

Comme nous l'avons vu précédemment, le test d'ILD suppose l'additivité des événements évolutifs entre les arbres. Ceci suppose que chaque caractère d'une matrice possède le même poids quelle que soit la matrice dans laquelle il se trouve. La plupart des méthodes de correction de distances donnent des poids différents à une substitution selon qu'elle se trouve dans un environnement contenant beaucoup ou peu de substitutions. Cette propriété pose un problème quant au critère d'additivité puisque la substitution aura un poids différent dans l'alignement d'origine, dans l'alignement concaténé et dans chacun des alignements produits aléatoirement. Nous avons donc choisi d'utiliser la divergence observée entre couples de séquences, qui permet de conserver la propriété d'additivité aux erreurs d'arrondis près. Il pourrait ensuite être intéressant d'étendre cette étude à des méthodes de distances permettant de conserver (ou de ne pas trop violer) cette propriété.

La longueur des arbres doit être exprimée en nombre de substitutions. Pour exprimer la taille des arbres de distances en ces termes, nous avons sommé la longueur des branches puis multiplié par le nombre de sites présents dans l'alignement. Cette méthode pose le problème des arrondis sur la taille des arbres et les valeurs d'ILD car contrairement à la méthode de parcimonie, les longueurs des branches ne sont pas exprimées en entiers. Nous avons considéré pour les simulations présentées ici que les valeurs d'ILD sont différentes de

l'ILD initiale (ILD_0) lorsqu'elles diffèrent de plus de 0,2 % de la somme des tailles des arbres initiaux.

Comme dans le test ILD, notre test (ILD-BIONJ) compare le nombre de pas dans les arbres les plus courts construits à partir des données séparées à ceux obtenus avec les données combinées et réparties aléatoirement en deux alignements de tailles identiques aux alignements de départ.

Pour chaque étape, l'indice d'ILD peut être calculé :

$$ILD = L - \Sigma L_i$$

Où L représente la taille de l'arbre le plus court construit à partir de la matrice concaténée et ΣL_i représente la somme des tailles des arbres. Pour chaque jeu de données simulées, 1000 ré-échantillonnages ont été effectués. L'hypothèse de congruence peut être rejetée au risque de 5% lorsque l'indice de départ ILD_0 est supérieur à 95 % des valeurs issues des ré-échantillonnages (ILD_r).

Pour tester les performances de notre méthode et pouvoir la comparer à l'ILD, nous avons utilisé le même protocole de test que Darlu et Lecointre (Darlu et Lecointre, 2002).

2.5.3 ***Simulations***

Pour reproduire les conditions de test de Darlu et Lecointre (2002), nous avons utilisé le programme PAML développé par Yang (1997) pour simuler chaque alignement de huit séquences nucléotidiques en faisant varier les paramètres suivants :

- La forme des arbres peut être symétrique (SYM) ou asymétrique (ASYM) (fig. 2.33).
- Les taux d'évolution peuvent suivre une horloge moléculaire (CER pour « Constant Evolutionary Rate ») ou bien varier d'un facteur 3 d'une branche à l'autre (VER pour « Variable Evolutionary Rate ») (fig 3.33).
- Les séquences simulées peuvent avoir une longueur de 100 nucléotides ou de 1000 nucléotides.

- Le taux d'évolution *s* peut prendre les valeurs 0,02 ; 0,1 ; 0,2 et 0,4. Un taux d'évolution de 0,02 représente de l'ordre de deux substitutions pour 100 sites par branche, tandis qu'un taux de 0,4 correspond à environ 40 substitutions.

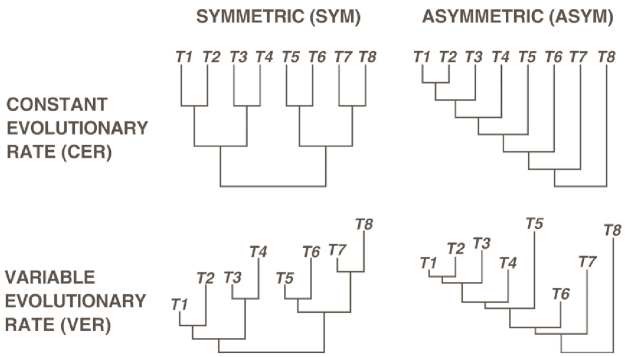


FIG. 1.—Symmetric (SYM) and asymmetric (ASYM) simulated trees according to constant (CER) or variable (VER) evolutionary rate. Long and short branch lengths are in the ratio of three when the evolutionary rate is variable (VER).

Fig. 2.33 : Les différents arbres sur lesquels sont simulées les séquences. Extrait de Darlu et Lecointre (2002).

- L'hétérogénéité des taux d'évolution entre sites peut être nulle (ce qui correspond à un paramètre de loi Gamma infini) ou varier selon une loi Gamma de paramètre α = 1,2 (hétérogénéité relativement faible), α = 0,6 (hétérogénéité moyenne) ou α = 0,06 (hétérogénéité extrême).

Chaque valeur présentée dans les tableaux correspond à 500 simulations indépendantes utilisant les mêmes paramètres de simulation.

Pour permettre une bonne comparaison entre les deux méthodes, nous avons parallèlement effectué les tests ILD (en utilisant PAUP* 4.0b10) et ILD-BIONJ sur les mêmes données simulées. Les résultats des tests ILD sont remarquablement semblables à ceux obtenus par Darlu et Lecointre (Darlu et Lecointre, 2002) sauf dans de rares cas (voir légende des tableaux).

2.5.4 ***Résultats et discussion***

Les tableaux 2.2, 2.3 et 2.4 montrent les résultats des tests dans différentes conditions où les données simulées sont congruentes entre elles, et permettent donc d'évaluer le risque de rejeter l'hypothèse de congruence des données lorsqu'elle est vraie (risque de première espèce). Le tableau 2.2 montre les résultats des tests ILD et ILD-BIONJ pour des paires d'alignements simulés dans des conditions identiques (topologies, hétérogénéité des taux d'évolution...). Dans ces conditions, les deux tests ne donnent des résultats erronés que dans moins de 5 % des cas quels que soient les paramètres utilisés. Ainsi, que les arbres soient symétriques ou non, que les taux d'évolution soient constants ou non, hétérogènes ou

homogènes et quelle que soit la longueur de l'alignement, les tests sont assez fiables si les données ont évolué de manière identique.

		HOM				HET α=0.6				HET α=0.06			
		SYM*SYM		ASYM*ASYM		SYM*SYM		ASYM*ASYM		SYM*SYM		ASYM*ASYM	
		L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000
CER	s= 0.02	(1.0) **2.8**	(0) **0**	(1.4) **3.0**	(0.6) **0.2**	(1.4) **2.4**	(2.8) **0**	(0.6) **4.4**	(3.4) **1.0**	(1.6) **3.4**	(0) **0**	(0.6) **4.6**	(2.0) **0**
	s= 0.10	(1.8) **0.4**	(0) **0**	(1.8) **1.6**	(0) **0**	(2.4) **4.6**	(2.8) **0.4**	(2.8) **4.8**	(3.0) **1.0**	(2.4) **1.2**	(0) **0**	(3.0) **2.0**	(2.4) **0**
	s = 0.20	(2.4) **0.4**	(0) **0**	(2.0) **1.4**	(0.8) **0**	(3.4) **3.0**	(2.8) **0.4**	(1.4) **2.6**	(2.8) **0.8**	(2.4) **0.4**	(0) **0**	(3.2) **2.2**	(3.4) **0**
	s = 0.40	(4.2) **0.4**	(0) **0**	(2.8) **1.0**	(4.2) **0**	(2.6) **3.4**	(5.2) **0**	(3.2) **3.0**	(4.4) **0.4**	(2.6) **2.0**	(0.2) **0**	(4.4) **3.6**	(3.8) **0**
VER	s = 0.02	(0.6) **2.0**	(0.4) **0**	(0.6) **1.6**	(0.8) **0**	(1.2) **2.4**	(1.6) **0.2**	(1.8) **2.8**	(2.8) **1.6**	(0.8) **2.8**	(1.0) **0.2**	(0.4) **1.4**	(2.2) **0**
	s = 0.10	(2.6) **0.8**	(0) **0**	(1.6) **1.8**	(1.0) **0**	(2.2) **4.0**	(4.4) **0.2**	(2.8) **3.0**	(3.4) **1.2**	(1.8) **1.2**	(0.4) **0**	(2.2) **3.4**	(2.8) **0**
	s = 0.20	(2.6) **0.2**	(0) **0**	(1.8) **1.2**	(2.6) **0**	(3.0) **6.2**	(2.6) **0.2**	(1.0) **2.8**	(4.2) **0.2**	(2.0) **0.6**	(1.2) **0**	(1.4) **2.6**	(4.2) **0**
	s = 0.40	(3.2) **0.4**	(0.2) **0**	(3.2) **0.6**	(4.6) **0**	(2.6) **4.2**	(3.4) **0**	(3.6) **3.8**	(2.8) **0.2**	(3.0) **1.2**	(2.6) **0**	(4.6) **1.6**	(3.0) **0**

Tableau 2.2 : Résultats des test ILD-BIONJ (en gras) et ILD (entre parenthèses) pour des alignements simulés dans des conditions identiques. Les chiffres indiquent la proportion (%) des simulations (n=500) conduisant à un rejet de l'hypothèse de congruence entre les données. Les résultats sont donnés en fonction des topologies (SYM=symétrique ; ASYM=asymétrique), des taux de substitution (s), dela variabilité des taux d'évolution entre branches (CER=taux constants ; VER=taux variables), de l'hétérogénéité entre site des taux de substitution (HOM=homogènes ; HET=heterogènes selon une loi gamma de paramètre α=0,6 ou α=0,06), et de la longueur des alignements simulés (L=100 ou L=1000).

Dans les tableaux 2.3 et 2.4, sont comparés les comportements des deux tests dans des conditions où les arbres sont congruents, mais où les conditions d'évolution des séquences sont différentes. Les deux tests montrent une assez bonne robustesse à la variation des taux d'évolution entre les branches des arbres comparés (tableau 2.3), avec peut-être un faible avantage au test ILD-BIONJ qui semble rejeter l'hypothèse de congruence des arbres moins souvent lorsque les séquences simulées sont courtes ou lorsque l'hétérogénéité entre les taux d'évolution est extrême (α=0.06). Il est à noter que, dans ce cas, malgré notre volonté de suivre le même protocole que Darlu et Lecointre (2002), nous n'avons pas réussi à reproduire les résultats obtenus par le test ILD dans quatres situations (marquées par un *). Dans ces conditions, Darlu et Lecointre (2002) trouvaient des taux de réponses erronées supérieurs à 10%.

		HOM		HET α=0.6		HET α=0.06	
		CER*VER		CER*VER		CER*VER	
		L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000
SYM	s= 0.02	(0.2) **1.8**	(0.2) **0**	(1.0) **2.6**	(0.4) **0**	(1.0) **3.6**	(4.4) **0.4**
	s= 0.10	(4.2) **1.4**	(0.0) **0**	(4.4) **2.8**	(0.4) **0**	(3.0) **5.4**	(4.8) **0**
	s = 0.20	(6.4) **0.8**	(0.0) **0**	(4.8) **2.0**	(0.4) **0**	(3.4) **4.0**	(5.2) **0.6**
	s = 0.40	(6.6) **1.6**	(0.2) **0**	(4.0) **0.8**	(2.2*) **0**	(2.2) **4.2**	(7.0) **0**
ASYM	s = 0.02	(1.8) **3.2**	(1.4) **0.2**	(1.0) **3.2**	(3.4) **0.8**	(1.2) **4.8**	(4.0) **4.0**
	s = 0.10	(4.0) **3.2**	(0.2*) **0.6**	(4.2) **4.0**	(3.4*) **1.4**	(2.2) **4.6**	(3.0) **1.2**
	s = 0.20	(5.6) **2.6**	(2.2*) **1.4**	(5.4) **2.6**	(6.6) **0.6**	(1.0) **2.4**	(5.6) **1.4**
	s = 0.40	(6.0) **3.2**	(12.6) **1.6**	(4.8) **4.2**	(7.0) **1.6**	(4.4) **4.0**	(7.4) **1.0**

Tableau 2.3 : Résultats des test ILD-BIONJ (en gras) et ILD (entre parenthèses) pour des alignements simulés dans des conditions identiques, sauf pour la constance des taux d'évolution entre branches de l'arbre (CER/VER). Les chiffres indiquent la proportion (%) des simulations (n=500) conduisant à un rejet de l'hypothèse de congruence entre les données. Les astérisques (*) montrent les cas où nos résultats et ceux de Darlu et Lecointre (2002) présentent de fortes différences pour l'ILD.

Les deux tests montrent également des taux d'erreur assez faibles lorsque l'on compare des jeux de données simulés en utilisant des paramètres de loi Gamma différents (tableau 2.4), sauf dans le cas où la différence est extrême (HOM*HETα=0.06). Dans ce cas, à la fois le test ILD et ILD-BIONJ sont incapables de prédire de manière fiable si des jeux de données de 100 nucléotides sont congruents. Cependant, lorsque les alignements simulés sont plus long (L = 1000), la prédiction est sensiblement améliorée pour le test ILD-BIONJ, contrairement à ce qu'on observe pour le test ILD.

		HOM* HET(α=0.6)				HOM* HET(α=0.06)			
		SYM		ASYM		SYM		ASYM	
		L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000
CER	s= 0.02	(1.0) **2.6**	(0) **0**	(0.6) **3.6**	(2.4) **0.4**	(4.2) **6.8**	(2.8) **0**	(3.8) **6.4**	(17.8) **3.6**
	s= 0.10	(4.0) **2.2**	(0) **0**	(3.0) **2.0**	(3.0) **0**	(47.0) **49.2**	(31.6) **0**	(20.8) **18.0**	(69.0) **11.0**
	s = 0.20	(3.4) **1.0**	(0) **0**	(4.0) **2.4**	(6.2) **0**	(37.5) **48.0**	(40.6) **0**	(15.6) **12.6**	(61.2) **4.0**
	s = 0.40	(3.4) **1.0**	(0) **0**	(3.6) **1.2**	(4.4) **0**	(9.2) **23.6**	(31.2) **0**	(6.8) **6.8**	(21.4) **0**
VER	s = 0.02	(0.8) **2.0**	(0.4) **0**	(1.2) **3.6**	(2.0) **0.2**	(4.8) **5.8**	(6.6) **0**	(4.4) **7.0**	(8.2) **2.0**
	s = 0.10	(3.6) **1.6**	(0.4) **0**	(5.0) **2.6**	(4.6) **0**	(43.8) **44.2**	(46.8) **0**	(33.4) **28.8**	(59.6) **4.4**
	s = 0.20	(4.0) **1.8**	(0.2) **0**	(4.0) **1.6**	(6.2) **0**	(35.2) **45.2**	(57.2) **0.2**	(23.8) **17.0**	(40.8) **0.6**
	s = 0.40	(3.0) **1.0**	(1.6) **0**	(3.0) **1.4**	(5.4) **0**	(8.8) **20.6**	(31.8) **0**	(5.4) **5.6**	(9.2) **0.4**

Tableau 2.4 : Résultats des test ILD-BIONJ (en gras) et ILD (entre parenthèses) pour des alignements simulés dans des conditions identiques, sauf pour l'hétérogénéité des taux d'évolution entre sites (HOM/HET). Les chiffres indiquent la proportion (%) des simulations (n=500) conduisant à un rejet de l'hypothèse de congruence entre les données.

Le dernier tableau (tableau 2.5) montre les résultats des tests pour la comparaison de données simulées à partir d'arbres différents. Ces comparaisons permettent donc d'avoir une idée de la tendance des tests à prédire que les données sont congruentes lorsqu'elles ne le sont pas (risque de deuxième espèce). Il est à noter que les arbres à partir desquels les données ont été simulées sont très différents ce qui constitue un cas d'incongruence particulièrement sévère. Même dans ce cas, l'on remarque que si les alignements sont courts, les taux de prédiction de l'incongruence des données sont relativement faibles et dépendent fortement de l'hétérogénéité des taux d'évolution et dans une moindre mesure de leur variabilité le long des branches. La situation est nettement améliorée si les alignements considérés sont plus long (L = 1000) et la variabilité des taux d'évolution n'est pas extrême.

		HOM		HET α=1.2		HET α=0.6		HET α=0.06	
		SYM*ASYM		SYM*ASYM		SYM*ASYM		SYM*ASYM	
		L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000	L = 100	L = 1000
CER	s= 0.02	(20.2) **34.2**	(100) **100**	(16.8) **28.6**	(100) **99.8**	(16.0) **29.4**	(100) **100**	(6.0) **14.8**	(92.4) **96.0**
	s= 0.10	(77.0) **84.0**	(100) **100**	(62.0) **78.0**	(100) **100**	(50.4) **67.0**	(100) **100**	(5.0) **7.6**	(60.2) **71.8**
	s = 0.20	(77.6) **91.6**	(100) **100**	(59.6) **79.4**	(100) **100**	(42.6) **62.2**	(100) **100**	(5.6) **8.0**	(51.4) **44.8**
	s = 0.40	(56.0) **78.2**	(100) **100**	(44.2) **68.0**	(100) **100**	(38.4) **53.8**	(100) **100**	(6.6) **9.6**	(61.8) **55.0**
VER	s = 0.02	(64.4) **71.6**	(100) **100**	(55.0) **67.0**	(100) **100**	(28.8) **60.2**	(100) **100**	(18.0) **32.6**	(100) **100**
	s = 0.10	(99.8) **99.6**	(100) **100**	(96.6) **97.6**	(100) **100**	(87.8) **92.4**	(100) **100**	(6.6) **13.0**	(83.0) **89.2**
	s = 0.20	(98.2) **99.8**	(100) **100**	(88.8) **95.4**	(100) **100**	(76.2) **87.0**	(100) **100**	(6.2) **10.6**	(83.0) **80.8**
	s = 0.40	(78.2) **94.4**	(100) **100**	(72.6) **84.8**	(100) **100**	(61.6) **74.2**	(100) **100**	(9.4) **12.4**	(93.4) **90.4**

Tableau 2.5 : Résultats des test ILD-BIONJ (en gras) et ILD (entre parenthèses) pour des alignements simulés dans des conditions identiques, sauf pour la topologie de l'arbre à partir duquel les données sont simulées. Les chiffres indiquent la proportion (%) des simulations (n=500) conduisant à un rejet de l'hypothèse de congruence entre les données.

Ainsi, à la fois les risques de premier et de second ordre sont très comparables pour les méthodes ILD et ILD-BIONJ. Les simulations suggèrent que les résultats de ces tests sont assez fiables si la longueur des séquences est suffisante et si l'hétérogénéité des taux d'évolution entre sites n'est pas trop grande. Le test ILD-BIONJ, semble globalement être légèrement moins sensible notamment au paramètre longueur des alignements puisque tout en montrant des risques de premier ordre comparables pour les alignements courts (tableaux 2.2, 2.3 et 2.4 ; L = 100), il rejette systématiquement plus souvent l'hypothèse de congruence sur les données simulées avec des arbres différents (tableau 2.5 ; L = 100).

Le test ILD-BIONJ pourrait donc être une bonne alternative au test ILD puisqu'il est beaucoup plus rapide et permet tout aussi efficacement de trouver les alignements congruents, tout en augmentant légèrement sa capacité à détecter les alignements incongruents. Le test

ILD-BIONJ permet donc d'envisager des tests à très grande échelle. Ceci pourrait être très utile dans une approche de détection des sites de recombinaison au sein des alignements, où certaines méthodes proposent de tester la congruence entre tous les fragments candidats d'un alignement partitionné (Zelwer, manuscrit en préparation).

La mise en pratique de ce test dans la problématique de la détection des gènes potentiellement utiles à la résolution de la phylogénie des bactéries est plus problématique. En effet, aux échelles de temps qu'il faut alors considérer, une méthode basée sur la divergence observée des séquences est probablement sujette à de nombreux artefacts. Ceci pose le problème d'ajouter à la méthode des corrections telles que celles utilisées classiquement par les méthodes de distances. Nous avons dit plus haut que la plupart des méthodes n'étaient probablement pas appropriées du fait qu'une substitution est considérée différemment selon le contexte dans lequel elle se trouve. Ce problème se pose particulièrement lors des ré-échantillonnages, ce qui peut complètement fausser le test. Il reste à imaginer une mesure de distance qui prenne en compte cette contrainte. On pourrait imaginer d'appliquer des modèles d'évolution du type de ceux utilisés en parcimonie comme la parcimonie sur les transversions ou n'importe quel mode de pondération des différentes substitutions. Cependant, ces méthodes étant également utilisables en parcimonie, il reste à prouver dans quelle mesure le test ILD-BIONJ serait dans ce cas un réel progrès par rapport à l'ILD.

Chapitre 3 : L’analyse intrinsèque des génomes

3 Chapitre 3 : L'analyse intrinsèque des génomes

3.1 Introduction : le gène dans le génome

La génomique est encore parfois définie comme la science qui permet de découvrir à grande échelle des gènes que la génétique échoue à trouver. Cependant, bien que cet aspect ne soit pas négligeable, l'objet de la génomique est, comme son nom l'indique, plus spécifiquement l'étude de niveaux d'organisation supérieurs à celui du gène. Bien plus qu'un chapelet de gènes, le chromosome, et particulièrement celui des procaryotes, est en effet une structure complexe qui nécessite d'être stockée dans un espace réduit, répliquée fidèlement, transcrite parfois de manière intensive au niveau de ses phases ouvertes de lecture, tout cela le plus souvent simultanément. Viennent s'ajouter à ces contraintes d'autres mécanismes comme la traduction, qui, chez les procaryotes, se fait de manière concomitante à la transcription, la recombinaison qui modifie l'ordre des gènes ou permet l'insertion de séquences étrangères etc... Le gène fait donc partie d'une structure intégrée dont il subit les contraintes et les pressions de sélection. Ainsi, l'étude de l'évolution du gène d'une part, et du génome d'autre part, sont difficilement dissociables.

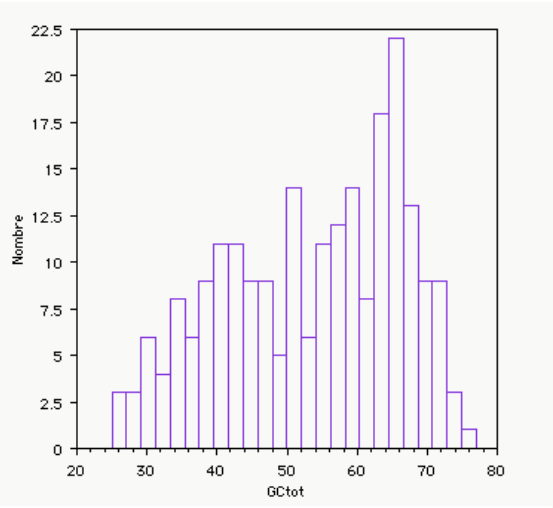


Fig. 3.1 : Distribution du taux de G+C global (GCtot) dans 224 genres bactériens (d'après les données de Galtier et Lobry, 1997)

Les procaryotes sont un modèle de diversité au niveau de la composition en bases du génome. Si chaque espèce montre une relative homogénéité du contenu en G+C de ses gènes, les comparaisons entre espèces révèlent une disparité impressionnante. Ainsi, il existe des mycoplasmes dont le contenu en G+C du génome est de 25 %, et des *Micrococcus* pour qui il peut atteindre jusqu'à 75 % (Fig. 3.1). Plusieurs hypothèses ont tenté d'expliquer ces variations sur la base de pressions de sélection, mais aucune ne semble avoir résisté à l'analyse des données.

Le taux de G+C des organismes procaryotes semble n'être corrélé ni à leur température optimale de croissance (Galtier et Lobry, 1997), ni à la vitesse de réplication du génome (Mira, *et al.*, 2001). Une possible relation avec la vie aerobie a été suggérée (Naya, *et al.*, 2002). Cette hétérogénéité est plus généralement interprétée comme le fruit de pressions de mutations directionnelles différentes (Sueoka, 1992), c'est-à-dire de biais de substitutions spécifiques de chaque espèce modéré par la sélection négative. Si elle est plus faible que cette variation interspécifique, l'hétérogénéité que montrent les gènes d'un génome est loin d'être négligeable (voir par exemple l'hétérogénéité du G+C3 d'*Escherichia coli* présenté fig. 1.11). Chacun des mécanismes mentionnés plus haut a un impact direct ou indirect sur l'évolution du génome et des gènes. Nous allons voir brièvement comment ces différents processus façonnent l'organisation, la composition et l'évolution des gènes et du génome. Nous nous attarderons un peu plus longuement sur le mécanisme de la réplication dont la connaissance sera utile pour la compréhension des résultats présentés.

3.1.1 ***La réplication***

Pour la majorité des bactéries connues, le génome se compose d'un chromosome unique et circulaire. Certaines peuvent posséder d'autres mini-chromosomes ou plasmides, la plupart du temps eux aussi circulaires. Cette « règle » souffre cependant de nombreuses exceptions comme par exemple dans le cas de la bactérie pathogène *Borrelia burgdorferi* dont le chromosome est linéaire et qui possède plusieurs plasmides qui peuvent être circulaires ou linéaires (Fraser, *et al.*, 1997). Nous allons nous intéresser plus spécifiquement au modèle de la réplication d'un chromosome circulaire bactérien, étudié notamment chez *E. coli* et *B. subtilis*.

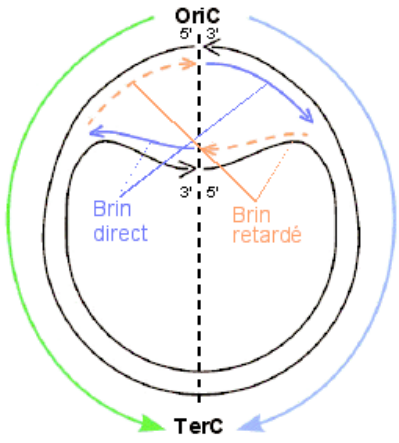


Fig. 3.2 : La réplication chez *E. coli*. La réplication s'initie au niveau de l'origine de réplication (OriC) de manière bidirectionnelle et se termine au niveau de la région du terminus. Le brin direct est répliqué de manière continue alors que le brin retardé est répliqué par fragments d'environ 2 kb (fragments d'Okazaki).

Le chromosome, chez *E. coli* et *B. subtilis* est répliqué de manière bidirectionnelle à partir d'une origine de réplication unique, et les chromosomes frères résultants sont ségrégés dans les deux moitiés

opposées de la cellule en division (fig. 3.2). La réplication des deux brins d'ADN est asymétrique : l'un des brins (le brin direct) est répliqué de manière continue et l'autre (le brin retardé) subit une réplication discontinue par petits fragments (les fragments d'Okazaki). Cette asymétrie est probablement à l'origine d'une différence de composition des deux brins que l'on retrouve chez un grand nombre d'espèces : les gènes codés sur le brin direct ont tendance à être plus riches en G qu'en C, et de manière un peu moins perceptible, plus riches en T qu'en A. Les raisons de ce biais sont encore discutées (Francino et Ochman, 1997; Frank et Lobry, 1999) bien que la plupart des auteurs s'accorde sur le fait que la désamination des cytosines méthylées peut jouer un rôle dans ce biais. Ce phénomène pourrait être lié à la réplication du génome ou/et à la transcription des gènes. Dans le cas d'un biais lié à la réplication, l'on pense en effet que le brin direct, qui sert de matrice à la réplication du brin retardé est plus souvent à l'état simple brin du fait de la réplication discontinue en fragments d'Okazaki et serait donc plus sensible à la transformation spontanée de la cytosine méthylée en thymine. Ce phénomène aurait pour conséquence de dépléter le brin direct en C et de l'enrichir en T. Cependant, ce modèle prédit un biais universel entre brin direct et brin retardé et il reste donc à expliquer pourquoi certaines bactéries, comme *Synecchocystis*, ne présentent pas de biais. L'asymétrie des brins n'affecte pas directement le taux de G+C des gènes, mais a un impact important sur leur usage du code et, de manière plus surprenante sur leur composition en acides aminés (Rocha, *et al.*, 1999b) ! Chez certaines espèces, comme *B. burgdorferi* et *Chlamydia trachomatis*, le biais résultant est si fort qu'il permet de prédire de manière presque certaine sur quel brin est codé un gène (McInerney, 1998).

Des observations récentes montrent que, contrairement à ce qui est souvent représenté, la machinerie de réplication ne se déplace pas le long du chromosome, mais que le chromosome passe à travers une « usine de réplication » (« replication factory ») fixée au niveau du plan de division de la cellule (Sawitzke et Austin, 2001). Selon ce modèle, réplication et ségrégation des chromosomes se font de manière concomitante. Les études de microscopie par fluorescence montrent que durant la réplication, les deux nouvelles origines de réplication s'éloignent rapidement chacune vers un pôle de la cellule (fig. 3.3A). Le moteur de ce tropisme est encore inconnu. Il a été proposé qu'un ensemble de protéines non identifiées pourraient jouer un rôle analogue aux protéines mitotiques (Sawitzke et Austin, 2001), mais d'autres auteurs proposent plus simplement que les mécanismes combinés de la transcription et de la traduction de protéines membranaires et de la translocation pourraient,

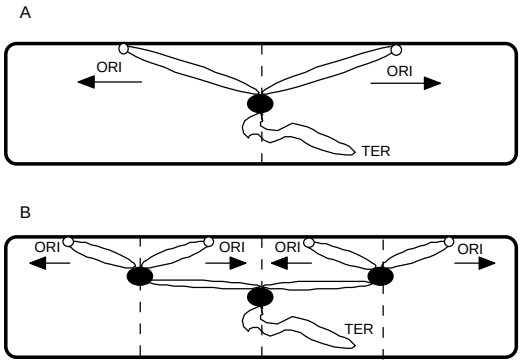


Fig. 3.3 : **A** : Le modèle de « replication factory » : Le complexe de réplication est fixé au plan de division de la cellule, et le chromosome passe à travers, les deux chromosomes nenosynthétisés sont ségrégés en même temps. **B** : Le modèle de réplication en oignon. D'après Sawitzke et Austin, 2001 modifié.

en ancrant le chromosome à la membrane, jouer ce rôle. Ce modèle a été nommé transertion (voir revue dans Woldringh, 2002).

Dans les cellules en phase de croissance exponentielle, il a été montré que l'origine de réplication pouvait se trouver présente en plusieurs copies, ce qui suggère que dans ces conditions, l'initiation d'une nouvelle phase de réplication n'attend pas la fin de la précédente (fig. 3.3B). Cette réplication en « oignon » semble s'accorder assez bien avec le modèle de la « replication factory » (Sawitzke et Austin, 2001; Woldringh, 2002).

Le modèle de la « replication factory » permet d'expliquer d'autres phénomènes dont celui des inversions symétriques par rapport à l'origine et au terminus de réplication, décrit par Eisen *et al.* (Eisen, *et al.*, 2000) (fig. 3.4). Lorsque l'on place sur un graphique la position d'un orthologue d'une espèce bactérienne en fonction de sa position dans une espèce proche, on obtient presque systématiquement un graphe en forme de X, dont le point d'intersection est l'origine ou le terminus de réplication (selon sa position sur les axes). Ce X révèle que la plupart des réarrangements dans les génomes bactériens se font de manière à ce que chaque gène conserve sa distance à l'origine et au terminus. La représentation des « dot-plots » d'*Escherichia coli* K12 avec les deux *Salmonella,* et des deux *Salmonella* entre elles illustre bien ce fait (fig. 3.5) : On constate une petite inversion autour de

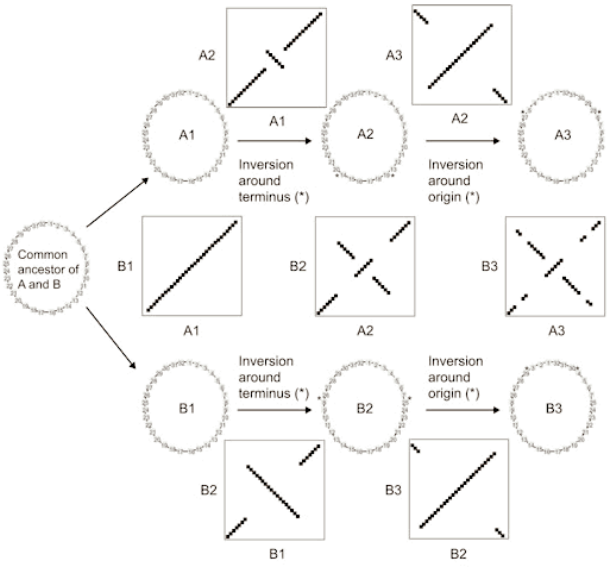


Fig. 3.4 : Modèle des inversions de génome. Le modèle montre un évènement de spéciation (entre A et B) suivi d'inversions autour de l'origine et du terminus. L'évolution du profile de dot-plot au fil de ces évènements est montré. Extrait de Eisen, *et al.*, 2000

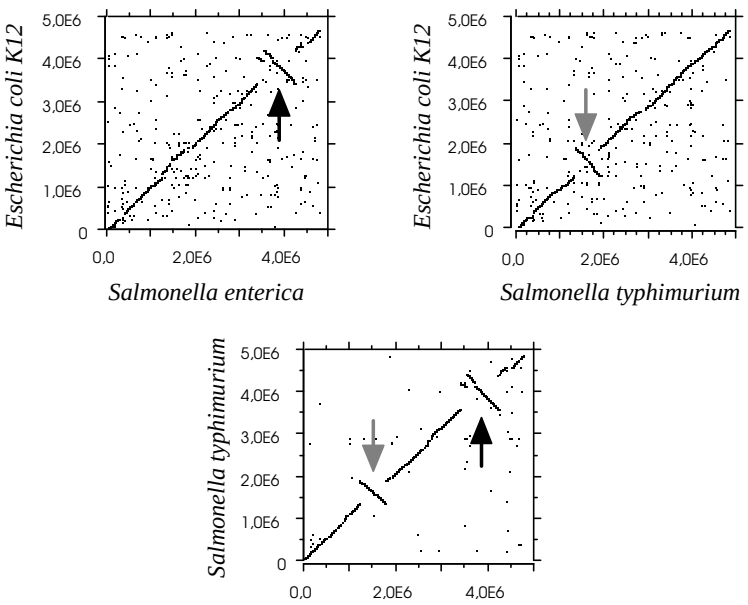


Fig. 3.5 : Dot-plot des génomes d'*E. coli*, *S. enterica* et *S. typhymurium*. Le groupe des *Salmonella* étant monophylétique, on peut déduire qu'une inversion a eu lieu autour du terminus dans la lignée de *S. enterica* (flèche noire) et qu'une autre a eu lieu autour de l'origine chez *S. typhimurium* (flèche claire).

l'origine de réplication entre *E. coli* et *S. enterica* et une autre inversion, cette fois autour du terminus entre *E. coli* et *S. typhimurium*. La comparaison des deux *Salmonella* permet de situer ces évènements dans le temps : aucune inversion n'a eu lieu dans la branche menant à *E. coli* et chacune des *Salmonella* a indépendamment subi une inversion.

Plusieurs explications ont été proposées dont certaines invoquent des avantages sélectifs à conserver uniquement ce type de réarrangements. S'il est vrai qu'une inversion n'incluant pas l'origine ou le terminus de réplication aura pour conséquence une interversion des brins directs et indirects et probablement des taux d'évolutions très forts pour les gènes concernés, la « replication factory » permet d'expliquer ce phénomène plus simplement par la proximité des deux fourches de réplication symétriques, qui favorise la probabilité d'une recombinaison non homologue.

La terminaison de la réplication se fait chez *E. coli* et *B. subtilis* au niveau d'un site placé approximativement à 180° de l'origine. Comparativement à l'initiation de la réplication, le phénomène de terminaison a été assez peu étudié, jusqu'à récemment où plusieurs travaux ont montré que réplication et division cellulaire se faisaient simultanément et où le rôle du terminus dans les étapes finales de la division a été suggéré (Perals, *et al.*, 2001; Capiaux, *et al.*, 2002). La réplication produit au mieux deux chromosomes entremêlés appelés « caténats » (« catenates » en anglais), au pire, un dimère de chromosomes (deux chromosomes liés de manière covalente) si un événement de recombinaison homologue a eu lieu entre les deux brins d'une même fourche (Lewis, 2001). La résolution des caténats se fait par l'action d'une topoisomérase (TopoIV chez *E. coli*). Le problème des dimères de

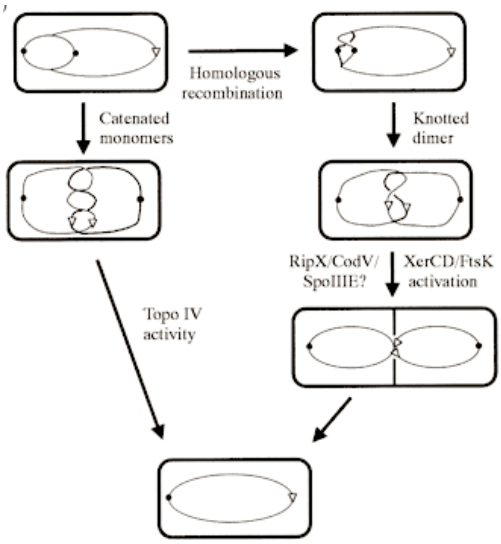


Fig. 3.6 : Résolutions des dimères (gauche) et des caténats (droite) de chromosomes chez *E. coli*. Voir détails dans le texte. Extrait de Lewis, 2001.

chromosomes est plus complexe et l'absence de résolution provoque de graves difficultés de ségrégation. L'action d'une recombinase (XerCD) dans la région du terminus et de la protéine FtsK (dont le rôle dans la division cellulaire est par ailleurs bien connu) est essentielle au bon déroulement de cette étape. Le site *dif* est un site d'action préférentielle de la topoisomérase TopoIV et est absolument nécessaire l'action de la recombinase XerCD (fig. 3.6). Il semble donc important que les fourches de réplication se rencontrent au niveau de ce site. Il existe, chez *E. coli* et *B. subtilis*, un ensemble de séquences appelées *ter* situées de part et d'autre de *dif* dont la fonction est de favoriser la rencontre des fourches de réplication à proximité de ce site. Les séquences *ter* constituent les sites de fixation de la protéine Tus (RTP chez *B. subtilis*), dont la fonction est d'empêcher le passage de l'hélicase précédant le complexe de réplication (DnaB), et ce de manière polaire, c'est-à-dire qu'elle laisse entrer les fourches dans la région du site *dif* sans entrave, mais bloque le passage d'une fourche s'éloignant de ce site (voir Fig. 3.7). Ainsi, si pour une raison ou une autre, l'une des fourches est en retard sur l'autre, cette dernière ne pourra pas dépasser la zone de terminaison. L'ensemble des sites *ter* forme donc, chez *E. coli* et *B. subtilis* un piège pour une fourche de réplication qui, une fois entrée dans cette zone ne peut ni avancer ni reculer. Cette fourche reste donc arrêtée jusqu'à l'arrivée de l'autre fourche et la terminaison de la réplication. Nous reviendrons un peu plus tard sur ce point.

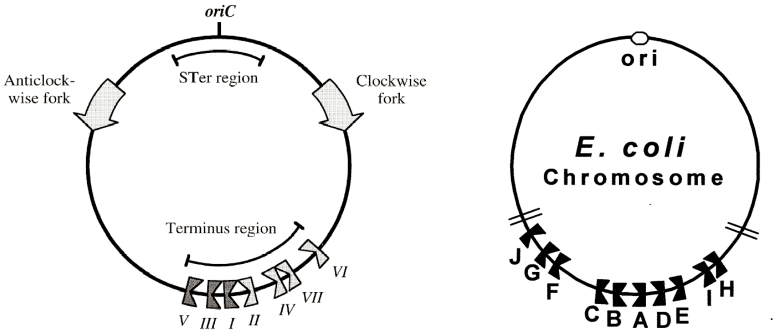


Fig 3.7 : Les chromosomes de *B. subtilis* (gauche) et *E. coli* (droite). L'origine ainsi que les différents sites *ter* connus sont représentés par des chiffres romains ou par des lettres. L'orientation des sites *ter* symbolise la polarité de leur action. Chez *B. subtilis* par exemple, les sites *VI, VII* et *IV* laissent passer la fourche progressant dans les sens des aiguilles d'une montre, mais les sites *I, III* et *V* la bloquent. Extrait de Wake, 1997 (pour *B. subtilis*) et Mulugu, *et al.*, 2001 (pour *E. coli*).

3.1.2 ***L'expression : transcription et traduction***

Comme je l'ai déjà souligné, la transcription des gènes en ARN messagers et la réplication se font simultanément, notamment pendant les phases de croissance. Il en résulte des contraintes supplémentaires quant à l'organisation du chromosome. Par exemple, le couplage transcription/traduction/translocation ancre le chromosome à la membrane au niveau des gènes codant pour des protéines membranaires ou excrétées, et le subdivise en une quarantaine de domaines indépendants du point de vue de leurs contraintes topologiques (Woldringh, *et al.*, 1995). Il est ainsi possible que ce type de facteur influence la disposition des gènes sur le chromosome. D'autres caractéristiques comme le taux d'expression peuvent façonner l'organisation du génome, comme chez *Bacillus subtilis* où les gènes fortement exprimés semblent être regroupés à proximité de l'origine de réplication (Kunst, *et al.*, 1997). L'explication peut en être un effet de dosage, du fait de la présence d'un plus grand nombre de copies des gènes liés à l'origine pendant la réplication, ou une adaptation à la compartimentation décrite dans le modèle de « replication factory ».

On a remarqué que chez la plupart des bactéries, les gènes ont tendance à être codés de manière prédominante sur le brin direct. Cela peut aller d'un léger excès, comme chez *E. coli* chez qui 55 % des gènes sont sur le brin direct, à un biais très important comme chez *B. subtilis* ou *Mycoplasma genitalium* (respectivement 75 % et 80 %) (Rocha, *et al.*, 1999b). Le biais d'orientation des gènes pourrait être du à une pression de sélection pour éviter les collisions frontales entre l'ADN- et l'ARN-polymérase. La transcription étant, à l'instar de la réplication, un processus asymétrique, il a été suggéré que l'asymétrie de composition des brins décrite plus haut pouvait également avoir son origine dans un biais mutationel lié à la transcription additionné au biais d'orientation des gènes (Francino et Ochman, 1997). Cependant, les gènes les plus fortement exprimés ne montrent pas d'asymétrie de composition plus marquée, ce qui semble en contradiction avec ce modèle (Rocha, *et al.*, 1999b).

L'usage des codons est fortement biaisé dans les gènes dont le taux d'expression est fort chez de nombreuses espèces bactériennes, et correspond aux ARN de transfert les plus abondants dans la cellule (Ikemura, 1981; Gouy et Gautier, 1982; Bulmer, 1987; Kanaya, *et al.*, 1999). Ceci suggère que ces gènes subissent une pression de sélection liée à la traduction, pour augmenter la probabilité de rencontrer l'ARNt correspondant aux codons à traduire. Cette adaptation peut s'interpréter soit en terme de rapidité de la traduction (le ribosome passe

moins de temps à attendre le bon ARNt s'il est abondant), soit en terme de fidélité (le ribosome chargera moins souvent un acide aminé erroné s'il trouve rapidement l'ARNt approprié). Ainsi, l'existence de cette sélection traductionelle suggère l'existence de deux classes de gènes, l'une dont l'utilisation des codons serait principalement déterminée par les biais mutationnels spécifiques de l'organisme considéré (les gènes peu exprimés), et l'autre dont l'usage du code serait contraint par la disponibilité des ARNt (les gènes fortement exprimés) (Ikemura, 1981; Gouy et Gautier, 1982). Kurland et collaborateurs (Andersson et Kurland, 1990; Berg et Kurland, 1997) ont montré que cette dernière classe de gènes correspondait plus spécifiquement aux gènes fortement exprimés pendant la phase exponentielle de croissance. D'autre part, Lobry et Gautier (Lobry et Gautier, 1994) ont montré que pour certains gènes fortement exprimés, même l'usage des acides aminés pouvait être adapté aux ARNt les plus fréquents dans la cellule. Cependant, il a été montré que de nombreux facteurs autres que les biais mutationnels et la sélection traductionnelle telle que nous l'avons définie avaient une influence sur l'usage du code des gènes. Ainsi, il existe un lien entre la structure secondaire d'une protéine et l'usage du code du gène correspondant, les protéines ayant un processus de repliement lent semblant utiliser préférentiellement des codons rares (Thanaraj et Argos, 1996a; Thanaraj et Argos, 1996b) et le biais d'usage du code varie entre le début et la fin d'un gène (Eyre-Walker et Bulmer, 1993).

3.1.3 ***Autres contraintes***

Les génomes diffèrent également au niveau de leur composition en mots (c'est-à-dire en oligonucléotides). Karlin *et al.*, (Karlin et Burge, 1995; Karlin et Mrazek, 1997; Karlin, *et al.*, 1997; Karlin, 1998; Karlin, *et al.*, 1998; Karlin, 2001) ont par exemple montré que chaque génome pouvait être caractérisé par une signature liée à la fréquence des différents oligonucléotides possibles. Par exemple, le dinucleotide CG est plutôt sur-représenté chez les protéobactéries des groupes α et β, mais est fortement sous-représenté chez la plupart des autres bactéries et archées. Les raisons de ces différences sont encore mal connues. Karlin *et al.* (Karlin, *et al.*, 1997) proposent qu'il existe des biais de mutations contextuels différents dans les espèces ou encore des contraintes structurales liées à la flexibilité des différents dinucléotides. De manière intéressante, la discrimination entre espèces se fait d'autant mieux que les mots considérés sont longs. Certains mots longs et notamment les mots

palindromiques sont évités dans de nombreux génomes, probablement du fait que ce type de mots est souvent la cible des enzymes de restriction (Rocha, *et al.*, 2001).

Une autre caractéristique importante des génomes est leur contenu en séquences répétées ne correspondant pas à des gènes dupliqués. En effet, les répétitions de séquences de plus de 25 nucléotides sont très fortement sur-représentées dans certains génomes bactériens et archéens par rapport à ce que l'on attend par hasard (Rocha, *et al.*, 1999a; Achaz, *et al.*, 2002). Ces répétitions, formant des sites potentiels de recombinaison intra-chromosomique ont un impact fort sur la dynamique du génome. La recombinaison entre répétitions directes provoque des délétions ou des duplications des régions présentes entre les répétitions alors que la recombinaison entre répétitions inversées conduit à des inversions. Certains génomes, comme par exemple chez *Mycoplasma genitalium* et *M. pneumoniae*, possèdent un forte densité de répétitions dans certaines régions du génome et il a été proposé qu'elles pourraient constituer un mécanisme de production de nouvelles formes de protéines de surface pour échapper aux défenses immunitaires de l'hôte (Rocha et Blanchard, 2002). Dans le même ordre d'idées, les gènes impliqués dans la réponse au stress chez *E. coli* présentent des densités importantes de répétitions courtes en tandem qui pourraient permettre (ou être la trace) des événements de formation de nouveaux allèles dans des environnements changeants (Rocha, *et al.*, 2002).

3.2 La structuration du GC3 et des taux d'évolution.

Comme nous venons de le voir, la séquence d'un gène est contrainte à différents niveaux, qui peuvent être directement ou indirectement liés à sa fonction (produit, expression) ou bien résulter de contraintes à des niveaux d'organisation supérieur au gène (réplication, mutation et maintien de l'intégrité du chromosome). Cependant, la superposition des différents biais fait qu'ils peuvent être difficiles à identifier chez certains organismes. Il n'est pas exclus que d'autres biais encore ignorés puissent jouer un rôle. Par exemple, Sharp *et al.* (Sharp, *et al.*, 1989) ont montré que le taux d'évolution des gènes à proximité de l'origine de réplication était approximativement deux fois moins élevé que pour les gènes à proximité du terminus de réplication. Ceci suggère que la localisation sur le chromosome est une source potentielle de différences dans le processus mutationnel affectant les gènes, et donc une éventuelle source de biais de composition.

Un certain nombre d'études (Deschavanne et Filipski, 1995; Guindon et Perriere, 2001) suggèrent que chez *E. coli*, la composition des gènes en troisième position des codons varie en relation avec la proximité du terminus de réplication. J'ai donc complété et étendu cette analyse à 48 génomes bactériens et 11 génomes archéens.

3.2.1 ***Matériel et Méthodes***

3.2.1.1 Calcul des courbes de valeurs cumulées.

Les génomes complets et leurs annotations ont été extraits de la base de données EMGLib (Perriere, *et al.*, 2000a). Après sélection des séquences codantes contenant plus de 150 nucléotides, nous avons calculé l'indice d'adaptation du code (CAI) (Sharp et Li, 1987) et la fréquence de nucléotides G+C en troisième position des codons (G+C3) pour chacun des gènes. Le calcul du CAI se fait avec la formule suivante :

$$\ln(\mathrm{CAI}) = \sum_{i=1}^{61} f_i \ln w_i$$

Où f_i est la fréquence relative du codon *i* dans le gène et w_i est le rapport entre la fréquencee du codon *i* et la fréquence du codon synonyme majeur pour l'acide aminé considéré, ce rapport ayant été estimé dans un ensemble de gènes de réference.

Comme le calcul du CAI nécessite une table de référence, nous avons choisi de baser cet indice sur des gènes hautement exprimés. Nous avons donc à chaque fois utilisé les protéines ribosomiques comme référence car ces gènes sont connus pour être fortement exprimés chez les organismes unicellulaires comme les procaryotes (Srivastava et Schlessinger, 1990). Un indice de CAI élevé indiquera ainsi une grande richesse du gène en codons optimaux. Pour chacun de ces gènes, nous avons ensuite calculé les valeurs de ces paramètres centrées sur la moyenne (que nous noterons CAIc et G+C3c) afin de tracer la somme cumulée de ces valeurs le long du génome. Cette méthode permet d'intégrer et donc d'amplifier fortement les variations de ces paramètres le long du génome. Ainsi, une portion

de la courbe présentant par exemple une pente positive témoignera d'une zone du génome relativement homogène où le paramètre considéré est supérieur à la moyenne.

Lorsque la position de l'origine et du terminus de réplication ne sont pas disponibles dans les annotations du génome complet, nous les avons déterminées en utilisant le programme Oriloc (Frank et Lobry, 2000).

L'amplitude de la courbe des valeurs cumulées dépend de plusieurs facteurs dont essentiellement le nombre de gènes, la variance autour de la moyenne et la structuration des valeurs du paramètre considéré. Pour un génome donné, l'amplitude de la courbe est maximale si les gènes sont complètement ordonnés. Pour tester le significativité de cette structuration, nous avons simulé 1000 ordres de gènes aléatoires et ainsi déduit une distribution de l'amplitude attendue sous l'hypothèse d'une répartition aléatoire des gènes le long des chromosomes. La valeur de l'amplitude observée en comparaison de cette distribution permet de connaître le niveau de significativité de la structuration observée.

De même, nous avons testé pour certains génomes la structuration des taux de G+C des régions intergéniques.

3.2.1.2 Calcul de la divergence entre séquences.

Les indices de Ks (taux de substitution synonyme) et de Ka (taux de substitution non synonyme) (Li, *et al.*, 1985) ont été calculés pour l'ensemble des gènes ayant conservé leur position relativement à l'origine et au terminus de réplication dans plusieurs paires de génomes. Les gènes homologues entre paires de génomes proches ont été identifiés en utilisant le logiciel BLASTP2 (Altschul, *et al.*, 1997). Seules les protéines montant une E-value inférieure à 10^{-20} sont considérées. Les séquences nucléiques correspondantes sont ensuite alignées en fonction de l'alignement protéique de manière à conserver le cadre de lecture dans la comparaison des codons. Le Ka et le Ks sont ensuite calculés en utilisant les programmes JaDis (Goncalves, *et al.*, 1999) et PAML (Yang, 1997) pour vérification.

Pour étudier l'effet de la position d'un gène par rapport à l'origine et au terminus de réplication sur son taux d'évolution, il est important de ne prendre en compte que des gènes ayant conservé leur position. Pour ce faire, nous avons utilisé la technique de dot-plot décrite par Eisen *et al.* (Eisen, *et al.*, 2000), qui consiste à montrer sur un graphique la position des gènes considérés comme homologues dans les deux génomes (voir Fig. 3.4). Il existe plusieurs paires d'espèces proches pour lesquelles les génomes ont été complètement séquencés, par exemple chez les enterobactéries, et dans les genres *Listeria, Neisseria, Rickettsia, Helicobacter, Chlamydia, Mycobacterium, Mycoplasma, Streptococcus, Pyrococcus, Sulfolobus et Thermoplasma.* Cependant, dans certaines de ces paires de génomes, l'ordre des gènes est si mal conservé qu'il est impossible qu'il existe une conservation de la distance à l'origine depuis l'ancêtre commun. Un cas limite est constitué par la comparaison des génomes de *Streptococcus pneumoniae* et *S. pyogenes* présentée fig. 3.8. Si la courbe de correspondance des positions des gènes dans ces deux génomes forme une croix encore visible autour du terminus de réplication, celle-ci est si brouillée qu'il est difficile d'établir un critère de choix pour les gènes ayant effectivement conservé leur position au sein des deux génomes. Pour certaines paires de génomes, il est impossible de considérer que l'ordre des gènes a gardé quoique ce soit de l'ancêtre commun aux deux espèces. C'est le cas notamment *Mycobacterium tuberculosis* et *M. leprae,* et *Sulfolobus solfataricus* et *S. tokodaii*.

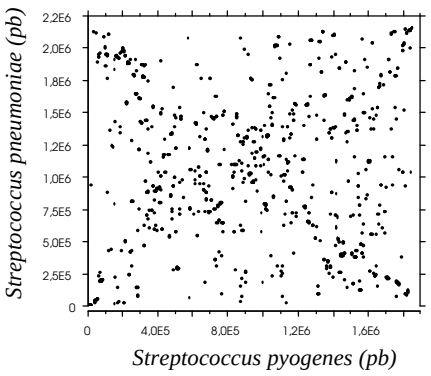
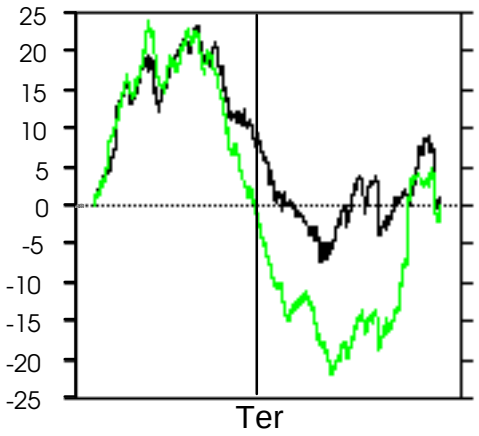


Fig. 3.8 : Dot-plot du génome de *Streptococcus pneumoniae* contre celui de *S. pyogenes*. Le terminus de replication est au centre du graphe (l'origine du graphe correspond à l'origine de réplication). Un très grand nombre d'inversions se sont produites.

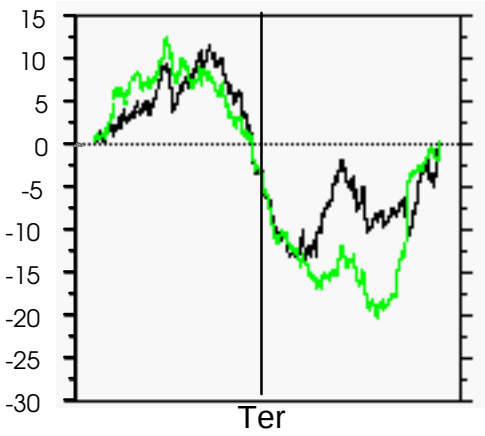
A l'inverse, parmi les autres génomes disponibles, certains sont si proches phylogénétiquement que leurs gènes sont presque complètement identiques au niveau nucléique, ce qui pose problème pour l'analyse statistique. C'est le cas notamment des deux souches de *Streptococcus pneumoniae* et de *Escherichia coli* O157:H7. Seulement huit paires restent donc exploitables pour cette analyse : *Salmonella*/*Escherichia*, *Escherichia* K12/*Escherichia* O157:H7, *Listeria monocytogenes/L. innocua*, *Neisseria meningitidis* souche A/*N. meningitidis* souche B, *Rickettsia prowazekii/R.conorii*, *Helicobacter pylori* J99/*H. pylori* 26695, *Chlamydia trachomatis/ C. muridarum* et *Pyrococcus abyssi/P. horikoshi*. Pour la paire *Salmonella*/*Escherichia*, il existe de nombreuses possibilités puisque trois souches d'*E.*

coli et deux souches de *Salmonella* sont disponibles. Cependant, les résultats obtenus sont identiques quelle que soit la paire considérée.

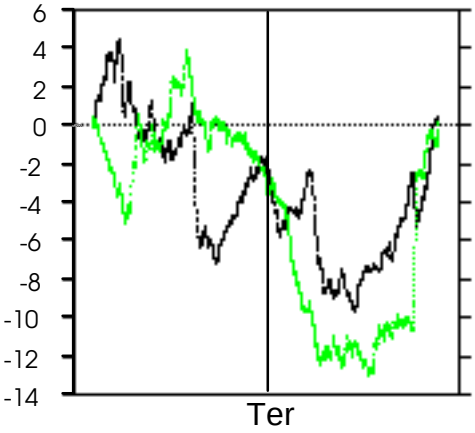
Escherichia coli O157:H7



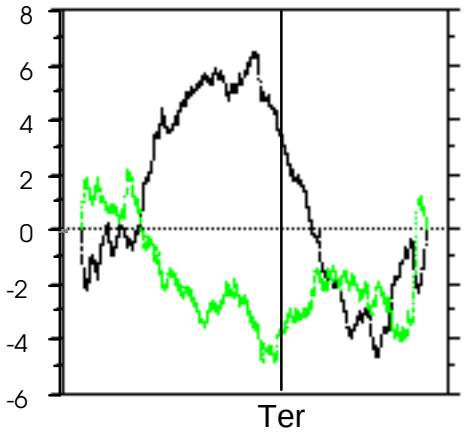
Salmonella typhimurium



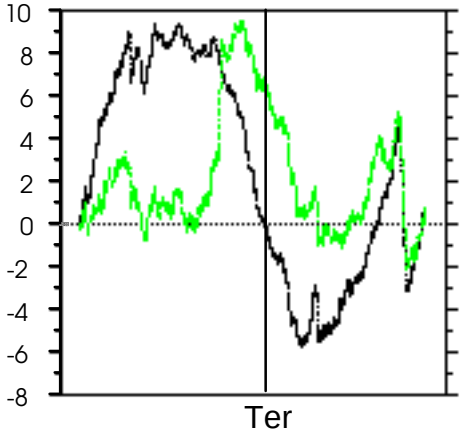
Vibrio Cholerae



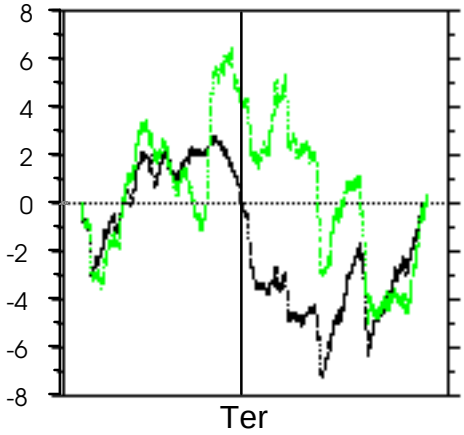
Pasteurella multocida



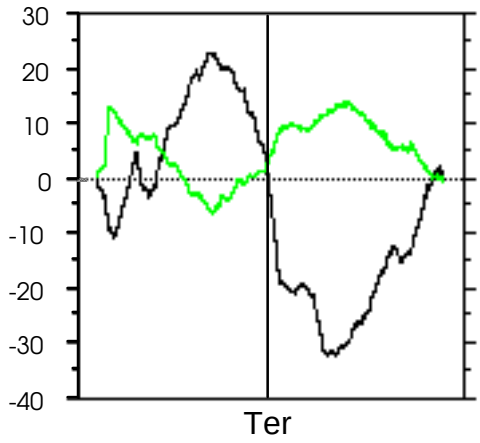
Sinorhizobium meliloti



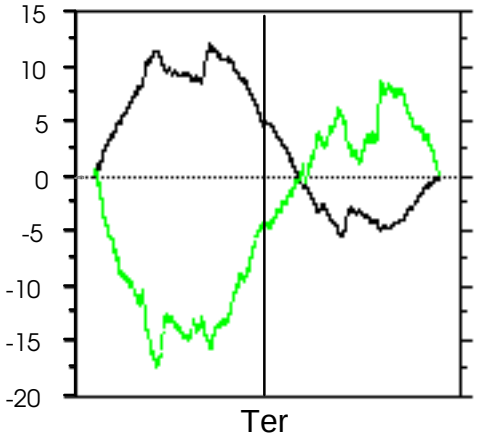
Brucella melitensis



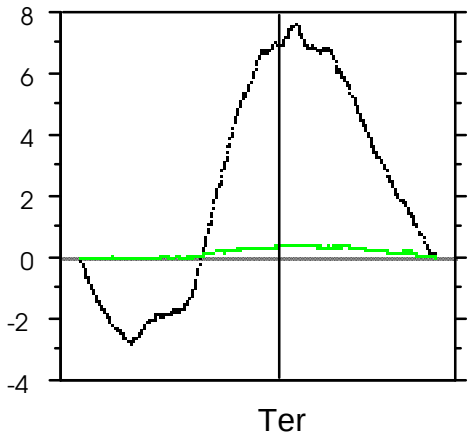
Bacillus subtilis



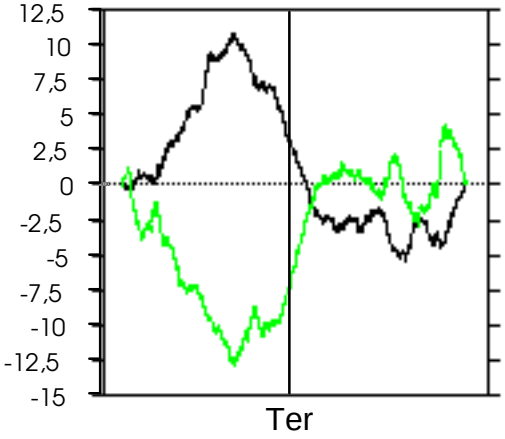
Staphylococcus aureus



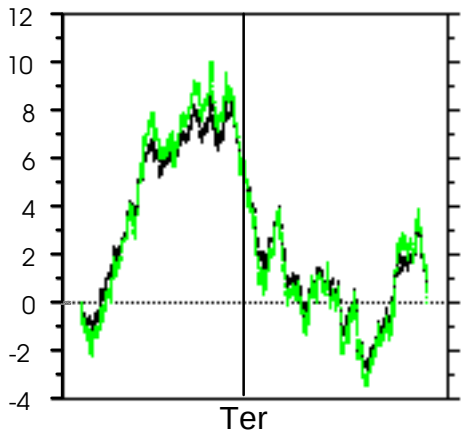
Mycoplasma genitalium

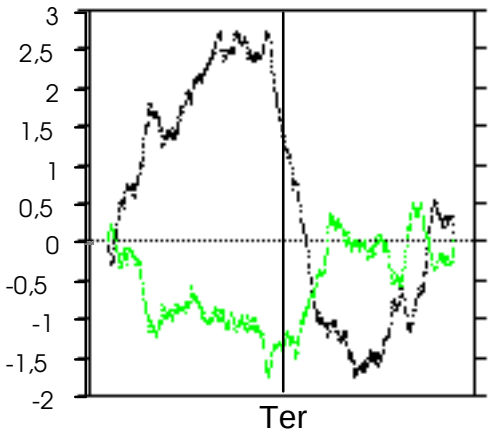


Listeria monocytogenes

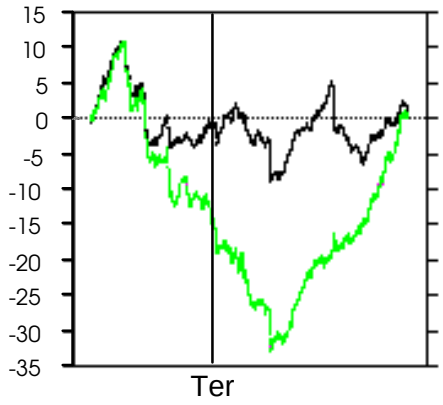


Mycobacterium tuberculosis

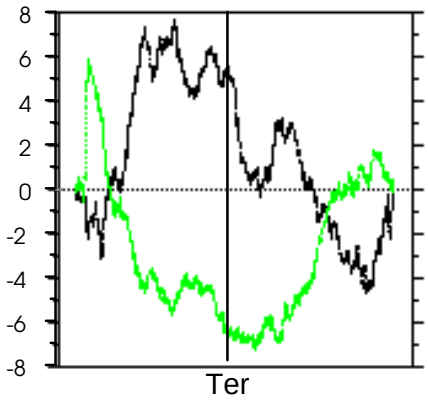




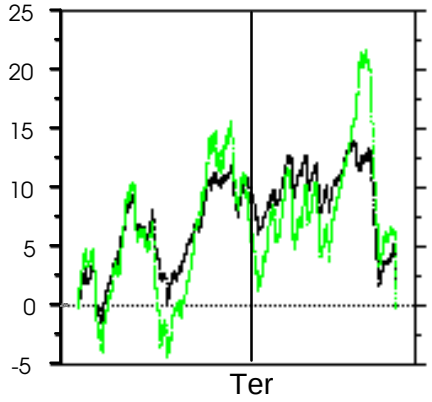
Pseudomonas aeruginosa



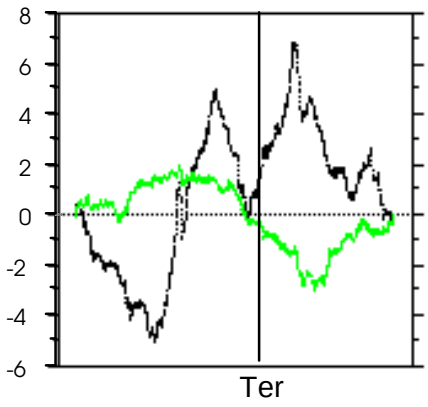
Bacillus halodurans



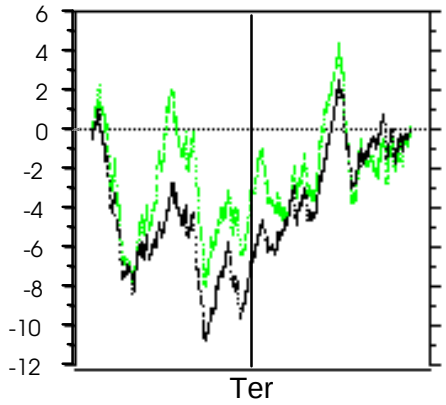
Ralstonia solanacearum Chr. 1



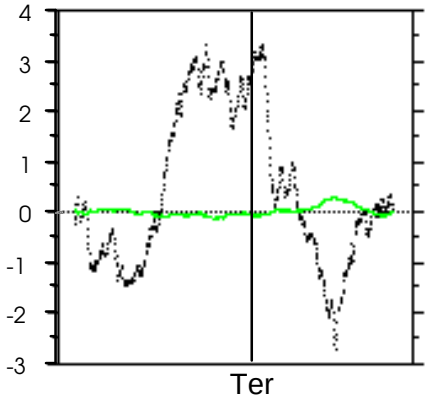
Thermotoga maritima

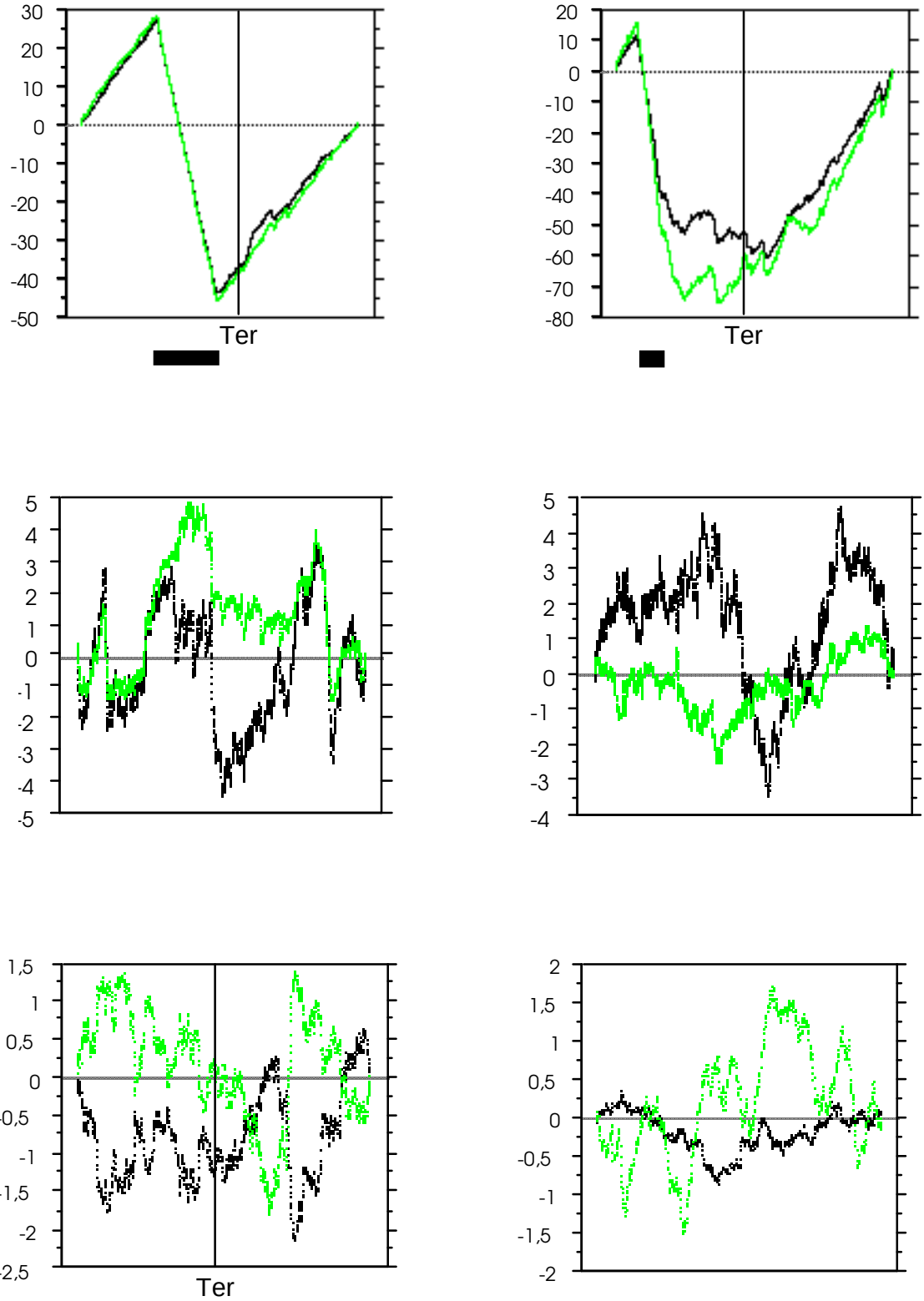


Ralstonia solanacearum Chr. 2

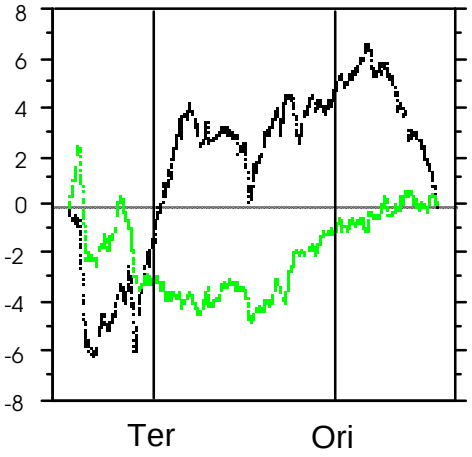


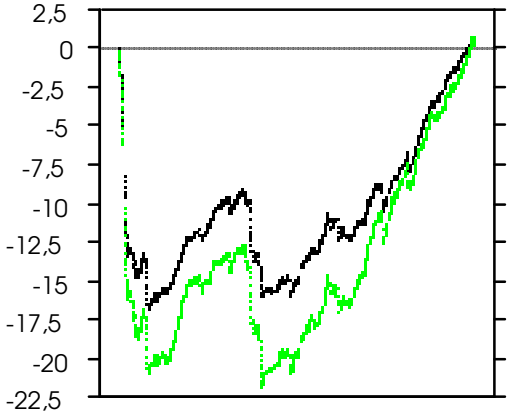
Mycoplasma pneumoniae

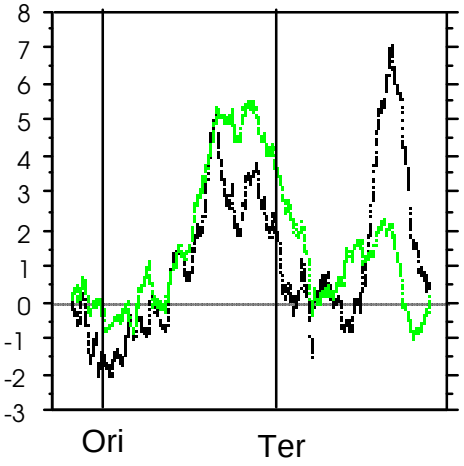


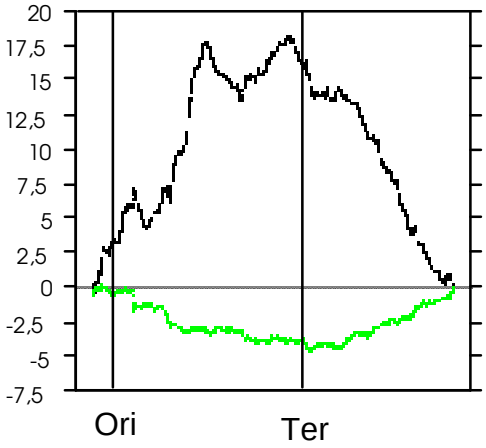


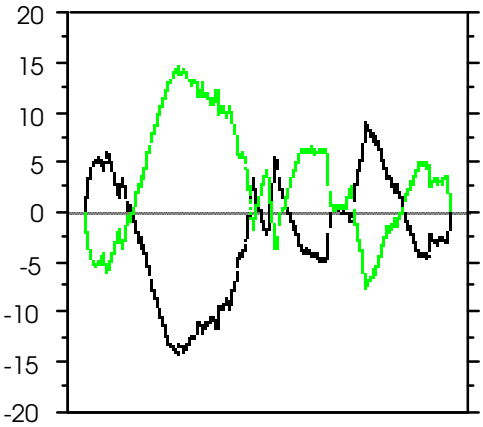
Pages précédentes : Fig. 3.9 : Graphes cumulés des valeurs centrées de G+C3 (G+C3c en noir) et de CAI (CAIc en clair) pour certains des génomes bactériens testés. L'origine de réplication, quand elle est connue, correspond généralement au point 0 en abscisse. Le terminus de réplication est indiqué par une barre verticale. Pour *Mesorhizobium* et *Deinococcus*, un rectangle noir représente les plasmides qui se sont probablement récemment insérés dans le chromosome. Les deux derniers génomes (*Rickettsia conorii* et *Borrelia burgdorferi*) ne sont pas significativement structurés.

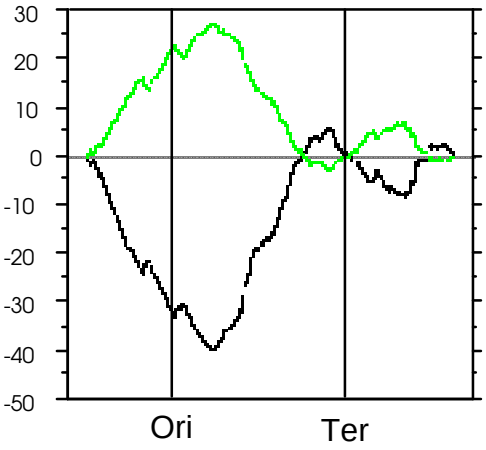

8
6
4
2
0
-2
-4
-6
-8
Ter
Ori

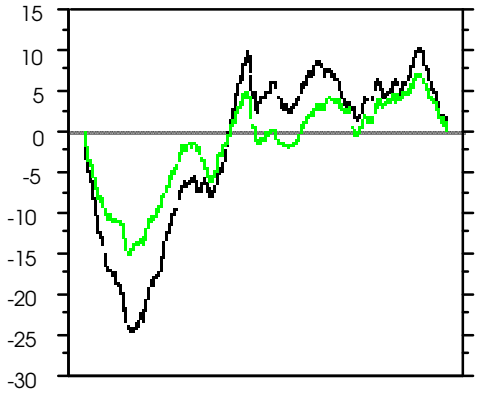

2,5
0
-2,5
-5
-7,5
-10
-12,5
-15
-17,5
-20
-22,5

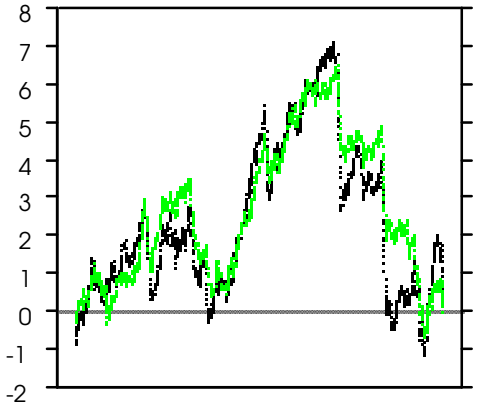

8
7
6
5
4
3
2
1
0
-1
-2
-3
Ori
Ter

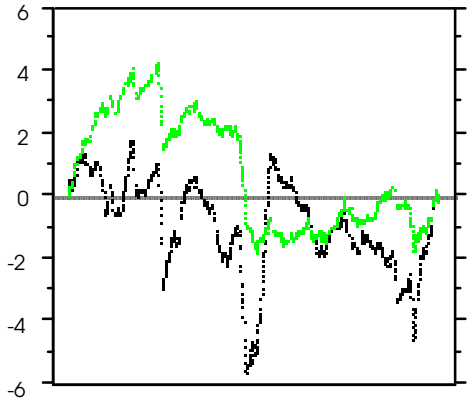

20
17,5
15
12,5
10
7,5
5
2,5
0
-2,5
-5
-7,5
Ori
Ter

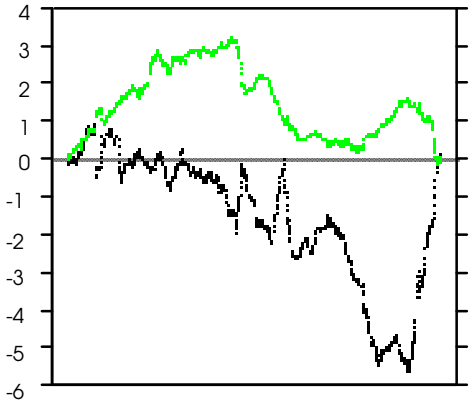

20
15
10
5
0
-5
-10
-15
-20

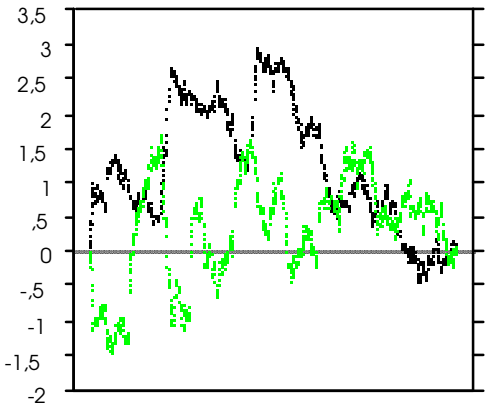

30
20
10
0
-10
-20
-30
-40
-50
Ori
Ter











Pages précédentes : Fig. 3.10 : Graphes cumulés des valeurs centrées de G+C3 (G+C3c en noir) et de CAI (CAIc en clair) pour les génomes archéens testés. L'origine de réplication, quand elle est connue, et par déduction le terminus de réplication (voir texte) sont indiqués par une barre verticale.

3.2.2 ***Résultats***

3.2.2.1 La structuration du taux de G+C en troisième position des codons

Les résultats de l'analyse de la structuration des taux de G+C3, CAI sont présentés dans le tableau 3.1 et les fig. 3.9 et 3.10. La plupart des génomes montrent une structuration fortement significative de leurs taux de G+C3 et du CAI (voir tableau 3.1). C'est le cas de 42 des génomes bactériens et de la totalité des génomes archéens analysés. Ainsi, d'une manière générale et contrairement à ce qui est supposé d'ordinaire chez les procaryotes, le contenu en G+C d'un gène n'est pas indépendant de sa position dans le génome. Ceci est particulièrement marquant pour certaines bactéries dont la structuration du taux de G+C semble très clairement être en relation avec la distance du gène à l'origine et au terminus. Il existe cependant de nombreux génomes pour lesquels la structuration ne suit pas un profil aussi clair et suggère la présence de nombreux fragments relativement homogènes au niveau de leur contenu en G+C. Un petit nombre d'espèces ne montre pas de structuration significative du taux de G+C le long de leur génome.

Étant donnée la fréquence des génomes bactériens montrant une structuration significative (44/48 à $p < 5.10^{-2}$), les quelques espèces non structurées (*Buchnera* sp. APS, *Rickettsia conorii*, *Borrelia burgdorferi* et *Aquifex aeolicus*, voir Tableau 3.1) apparaissent comme des exceptions. Il est intéressant de noter que trois d'entre elles (*Buchnera* sp. APS, *R. conorii* et *B. burgdorferi*) sont des parasites intracellulaires obligatoires, ce qui confère à leur génome un mode d'évolution très particulier du fait du relâchement d'un certain nombre de pressions adaptatives (Moran, 1996). De ce fait, ces bactéries ont un génome globalement riche en A+T et ont subi de fortes réductions de la taille de leur génome. Mais l'existence de très fortes structurations chez des espèces telle que les chlamydiales ou encore les *Mycoplasma*, parasites dont l'histoire est très analogue, interdisent d'y voir la seule explication pour une absence de structuration. On peut cependant concevoir que dans le processus aléatoire qui a conduit à la réduction de leurs génomes, ces bactéries aient perdu des gènes ou des fonctions qui sont responsables de la structuration chez d'autres espèces. Ainsi, chacune de ces espèces montreraient une absence de structuration pour des raisons indépendantes.

Phylum		Noms d'espèces	Gene #	$Mean_{G+C3}$	Str_{G+C3}	Str_{CAI}	Ter_{A+T}
Bacteria							
Gram positives	bas G+C	*Bacillus halodurans*	3950	42,1	+++	+++	n
		Bacillus subtilis	4052	43,6	+++	+++	y
		Staphylococcus aureus	2638	23,0	+++	+++	y
		Streptococcus pneumoniae	2015	35,6	+++	+++	y
		Streptococcus pyogenes	1682	31,7	++	+++	n
		Clostridium acetobutylicum	3651	21,3	+++	+++	y
		Lactococcus lactis subsp. lactis	2257	25,6	+++	+++	n
		Listeria innocua	2969	28,9	+++	+++	y
		Listeria monocytogenes	2849	30,0	+++	+++	y
		Mycoplasma genitalium	466	23,3	+++	+++	n
		Mycoplasma pneumoniae	674	41,1	+++	+	n
		Mycoplasma pulmonis	774	15,3	+++	++	n
		Ureaplasma parvum	607	12,9	+++	++	n
	haut G+C	*Mycobacterium leprae*	2691	49,6	++	+++	n
		Mycobacterium tuberculosis	4062	78,1	+++	+++	y
Cyanobacteria		*Nostoc* sp. PCC 7120	5329	35,2	+	-	?
		Synechocystis sp. PCC 6803	3103	49,6	+	+	?
Proteobacteria	γ	*Escherichia coli* O157:H7	5208	53,6	+++	+++	y
		Escherichia coli K12	4254	54,5	+++	+++	y
		Salmonella enterica	4519	56,2	+++	+++	y
		Salmonella typhimurium	4401	57,9	+++	++	y
		Buchnera sp. APS	562	14,4	-	-	-
		Vibrio cholerae	2562	48,6	+++	++	y
		Haemophilus influenzae	1647	29,0	+++	++	n
		Pseudomonas aeruginosa	5551	86,7	+++	++	n
		Xylella fastidiosa	2645	55,3	+++	++	y
		Yersinia pestis	3976	47,9	+++	++	y
		Pasteurella multocida	2011	34,4	+++	++	y
	β	*Neisseria meningitidis*	2065	58,7	+++	+	n
		Ralstonia solanacearum	3417	86,3	+++	+++	n
	α	*Sinorhizobium meliloti*	3326	78,8	+++	+++	y
		Mesorhizobium loti	6705	78,7	+++	+++	n
		Brucella melitensis	2055	65,9	+++	+++	y
		Agrobacterium tumefaciens	2679	71,6	+++	-	n
		Rickettsia conorii	1372	23,5	-	-	-
		Rickettsia prowazekii	830	18,4	++	+	y
		Caulobacter crescentus	3684	85,5	+++	+++	y
	ε	*Campylobacter jejuni*	1620	19,5	+++	+++	n
		Helicobacter pylori J99	1477	42,2	+++	+	n
		Helicobacter pylori	1513	41,5	+++	-	n
Chlamydiales		*Chlamydia muridarum*	797	33,5	+++	+	y
		Chlamydophila pneumoniae AR39	941	34,5	+++	+++	y
		Chlamydia trachomatis	891	34,6	+++	-	y
Spirochaetes		*Borrelia burgdorferi*	821	20,9	-	-	-
		Treponema pallidum	1000	54,8	++	+++	y
Thermotogales		*Thermotoga maritima*	1810	52,3	+++	+++	n
Aquificales		*Aquifex aeolicus*	1522	47,9	-	++	-
Deinococcales		*Deinococcus radiodurans*	2577	79,9	+++	++	n

Tableau 3.1 : voir légende page suivante.

Phylum	Noms d'espèces	Gene #	$Mean_{G+C3}$	Str_{G+C3}	Str_{CAI}	Ter_{A+T}
Archaea						
Crenarchaeota	*Aeropyrum pernix*	2694	65,3	+++	+++	?
	Sulfolobus solfataricus	2971	33,3	+++	+++	y (?)
	Sulfolobus tokodaii	2826	25,6	+++	+++	?
Euryarchaeota	*Halobacterium sp.* NRC-1	2017	87,3	+++	+++	?
	Methanococcus jannaschii	1674	27,7	++	+++	?
	Methanobacterium thermoautotrophicum	1859	55,9	+++	+++	n
	Pyrococcus abyssi	1764	50,2	+++	+++	y (?)
	Pyrococcus horikoshii	1979	42,9	+++	+++	y (?)
	Thermoplasma acidophilum	1477	54,1	+++	+++	?
	Thermoplasma volcanium	1495	41,0	+++	+++	?
	Archaeoglobus fulgidus	2374	57,8	+++	+++	?

Tableau 3.1 : Les espèces testées, leur appartenance phylogénétique (phylum), le nombre de gènes que contiennent leur chromosome (Gene #), le G+C3 moyen de leurs gènes ($Mean_{G+C3}$). Str_{G+C3} et Str_{CAI} montrent la significativité de la structuration observée respectivement pour le G+C3 et le CAI. (+++ : $p<10^{-3}$; ++ :$p<10^{-2}$; + :$p<5.10^{-2}$; - :non significatif). Ter_{A+T} indique si un enrichissement en A+T de la région terminus par rapport au reste du génome a été observée (y : enrichissement observé ; n : pas d'enrichissement ; ?: position du terminus incertaine ou inconnue).

Le cas particulier de *B. burgdorferi* suggère des explications très particulières à l'absence de structuration : cette bactérie possède en effet un chromosome linéaire et une très forte asymétrie des brins directs et indirects dont il a été montré qu'elle était le facteur majeur déterminant l'utilisation des codons chez cette espèce (McInerney, 1998). Chez cette bactérie, le processus de réplication du chromosome est donc très différent de la plupart des autres espèces bactériennes et doit subir des contraintes en conséquence.

Les deux espèces de cyanobactéries, *Nostoc* sp. PCC 7120 et *Synechocystis* sp. PCC 6803 ne montrent qu'une structuration faiblement significative ($p = 5.10^{-2}$). *Aquifex aeolicus* ne présente aucune structuration détectable. Le mécanisme de réplication chez ces bactéries est malheureusement très mal connu. Il est à noter cependant que chez ces bactéries, contrairement à la plupart des autres, il n'existe pas d'asymétrie de composition des brins liée à la position de l'origine et du terminus de réplication, ce qui interdit de les localiser en utilisant cette méthode (Karlin, *et al.*, 1998). Ceci suggère chez ces bactéries des mécanismes de réplication présentant des particularités qui pourrait expliquer leur absence de structuration.

Toutes les autres espèces de procaryotes analysées montrent une forte structuration du taux de G+C en troisième position. Ce phénomène semble donc être quasiment ubiquitaire,

puisque qu'il est absent de seulement deux grands phylum, celui des cyanobactéries (où la structuration est faiblement significative) et des aquificales. Cependant, le faible nombre d'espèces représentant ces phylum ne permet pas de généraliser cette absence de structuration. La structuration du G+C3 s'accompagne la plupart du temps d'une structuration du CAI. Cet indice peut être positivement ou négativement corrélé au G+C3 selon que les codons optimaux sont riches en G+C (par exemple *Mycobacterium tuberculosis,* fig. 3.9) ou en A+T (par exemple *Listeria monocytogenes,* fig. 3.9) respectivement.

On peut subdiviser les génomes structurés en deux grands types de profils approximativement aussi bien représentés l'un que l'autre :

- Le premier d'entre eux correspond à celui observé par Guindon et Perrière chez *E. coli* K12 (Guindon et Perriere, 2001). Il est caractérisé par un enrichissement en A+T de la région du terminus de réplication. On le retrouve chez des bactéries aussi diverses que *Pasteurella multocida*, *Sinorhizobium meliloti*, *Brucella melitentis*, *R. prowazeki*, *C. trachomatis*, *B. subtilis*, *Staphylococcus aureus*, *Listeria monocytogenes* et *Mycobacterium tuberculosis* et semble être représenté dans tous les grands phylums bactériens, à l'exception des cyanobactéries, des thermotogales et des aquificales.

- Le deuxième type de profil, tout aussi bien représenté correspond à une organisation mosaïque du génome dans laquelle un nombre assez important de régions relativement homogènes au niveau de leur taux de G+C se succèdent le long du génome.

Les archées présentent également de fortes structuration de leur taux de G+C. Le mécanisme de la réplication des archées commence tout juste à être élucidé, au moins pour certaines espèces. Il semble en effet que bien que les protéines impliquées dans la réplication de leur génome ont tendance à ressembler plus étroitement à des gènes eucaryotes, plusieurs d'entre elles possèdent un mécanisme de réplication de type bactérien, c'est-à-dire avec une origine unique d'où la réplication s'initie de manière bidirectionnelle (Lopez, *et al.*, 1999; Myllykallio, *et al.*, 2000; Zivanovic, *et al.*, 2002 ; MacNeill, 2001). Les positions des origines de réplication de *Pyrococcus horikoshii*, *P. abyssi*, *Methanobacterium thermoautotrophicum* et *Sulfolobus solfataricus* ont été prédites par des méthodes bioinformatiques (oligonucléotides-skew) (Lopez, *et al.*, 1999; She, *et al.*, 2001) et celle de *P. abyssi* a été confirmée expérimentalement (Myllykallio, *et al.*, 2000). L'origine de réplication semble

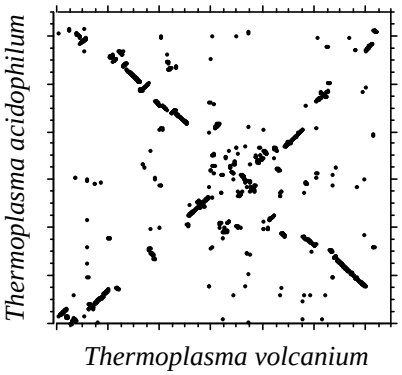


Fig 3.11 : Dot-plot du génome de *T. acidophilum* et *T. volcanium*. La croix suggère des inversions symétriques, comme couramment observé chez les bactéries. Le point central du graphe pourrait représenter le terminus (ou l'origine) de réplication.

coïncider avec le locus du gène *cdc6/orc1* qui est un homologue d'un gène impliqué dans la réplication eucaryote. Chez *Pyrococcus* et *Sulfolobus*, en supposant un terminus de réplication à 180° de l'origine, la région correspondante est effectivement A+T riche en comparaison de la moyenne du génome (Fig. 3.10.). Cependant, contrairement aux bactéries, cette région apparaît comme très courte et la localisation du terminus est peu sûre. Dans le cas de *Methanobacterium*, aucune région riche en A+T ne correspond au terminus inféré. Pour les autres archées considérées, avancer une position du terminus de réplication est encore plus risqué, puisqu'il a été suggéré par exemple que *Halobacterium* pourrait posséder plusieurs origines de réplication (Ng, *et al.*, 2000). De manière intéressante, le graphe représentant la position des homologues des deux espèces de *Thermoplasma* forme une croix qui suggère un mécanisme de réplication typiquement bactérien (Fig. 3.11). Ce type d'inversions se produisant de manière symétrique par rapport à l'origine et au terminus chez les bactéries, il est possible que le centre de cette croix représente soit l'origine, soit le terminus de réplication.

3.2.2.2 Variation des taux d'évolution le long du génome

Sharp *et al.* (Sharp, *et al.*, 1989) ont observé que les taux d'évolution synonymes (Ks) tendent à augmenter avec la distance des gènes à l'origine de réplication chez les entérobactéries. Puisque leur résultat reposait sur un jeu de données limitées, j'ai répété l'analyse en utilisant les génomes complets d'*E. coli* et de *S. typhimurium.* J'ai également analysé d'autres paires de génomes complets proches. Les résultats sont présentés dans la fig. 3.12. Dans trois des sept paires de génomes analysées, le Ks augmente avec la distance à l'origine de réplication. C'est également le cas pour les taux d'évolution non synonymes (Ka) chez les enterobactéries (*Salmonella/Escherichia*) et chez les chlamydiales (fig. 3.13). Chez *Chlamydia, Neisseria, Helicobacter* et *Pyrococcus*, aucune relation significative n'est trouvée entre le Ks et la distance à l'origine. Il est intéressant de noter que les paires pour lesquelles il existe un enrichissement en A+T marqué de la région du terminus de réplication (c'est-à-dire

les enterobactéries, *Rickettsia*, *Listeria* et *Chlamydia*) présentent également une structuration d'au moins un des deux taux d'évolution utilisés (Ks et Ka) ce qui peut suggérer une relation entre les deux phénomènes. Cependant, on peut remarquer que dans le cas des *Helicobacter* et des *Neisseria*, ce sont des souches d'une même espèce qui sont comparées et que les taux d'évolution sont faibles. Comme le suggère la comparaison d'*E. coli* K12 avec *E. coli* O157:H7 d'une part et avec *S. typhimurium* d'autre part, l'effet de la distance à l'origine pourrait bien n'être visible qu'après un temps de divergence suffisant. On ne peut donc exclure que l'effet de la distance au terminus ne devienne significatif chez *Neisseria* et *Helicobacter* après un temps de divergence suffisant. On peut remarquer dans le même ordre d'idée que l'augmentation du Ka n'est visible que dans les deux paires ayant le plus divergé (les entérobactéries et les chlamydiales).

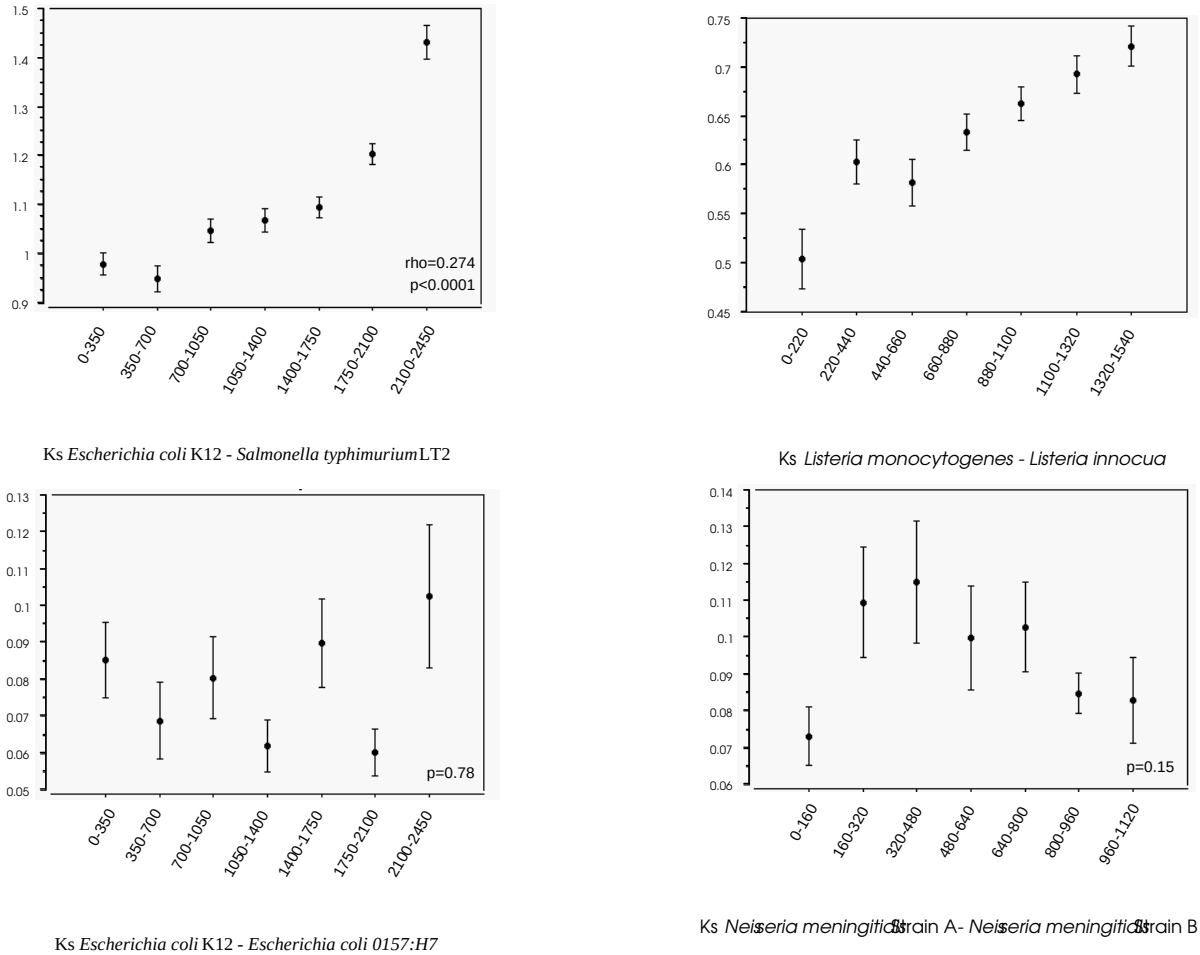
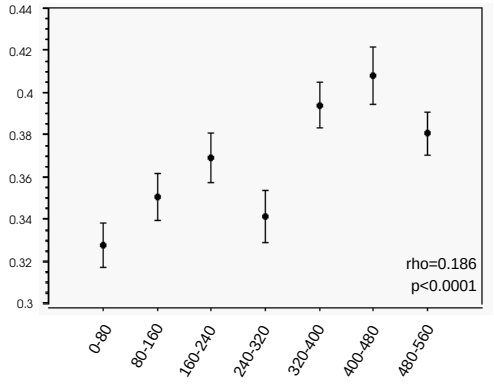
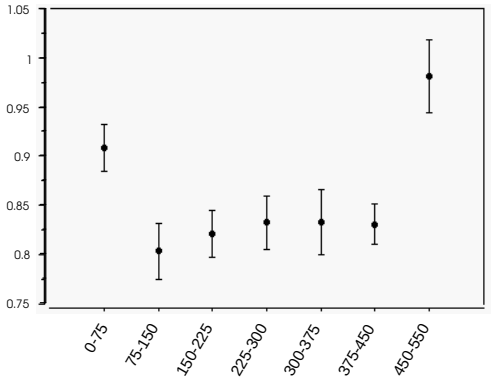


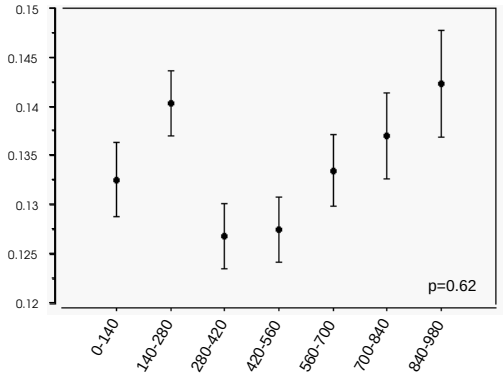
Fig. 3.12 :Taux d'évolution synonyme (Ks) en fonction de la distance à l'origine de réplication pour différents couples de génomes proches. Chaque génome a été divisé en 7 parts égales en fonction de l'éloignement des gènes à l'origine de réplication. Les intervalles de distances sont indiqués en kb. L'augmentation du taux d'évolution avec la distance à l'origine est testée avec le test des rangs de Spearman (rho et p indiqués dans chaque cadre). Les barres représentent 95 % d'intervalle de confiance.



Ks *Rickettsia prowazekii* - *Rickettsia conorii*



Ks *Chlamydia trachomatis* - *Chlamydia muridarum*



Ks *Helicobacter pylori* J99 - *Helicobacter pylori* 26695

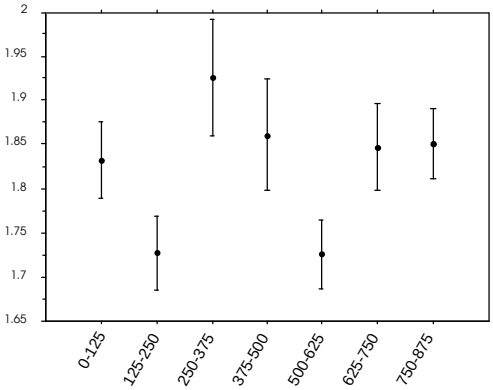
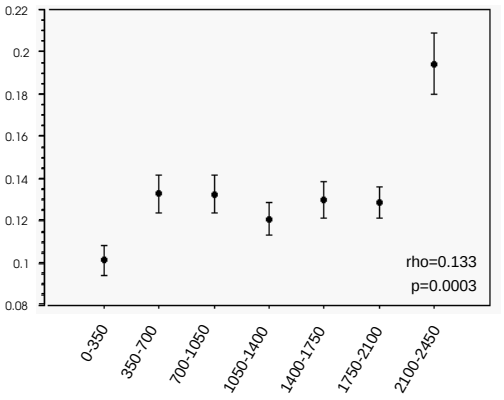
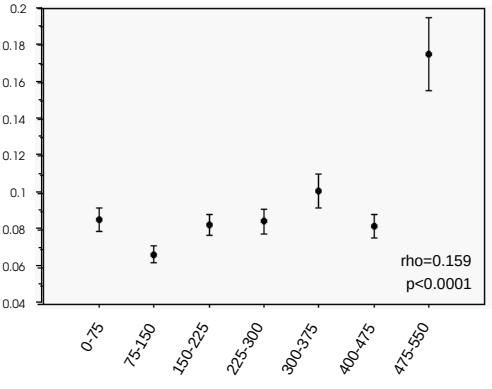


Fig 1 - suite 2



Ka *Escherichia coli* K12 - *Salmonella typhimurium* LT2



Ka *Chlamydia trachomatis* - *Chlamydia muridarum*

Fig. 3.12 suite : voir légende de la page précédente. Les deux dernières figures repésentent les taux d'évolution non-synonyme (Ka) pour les couples montrant une augmentation avec la distance à l'origine.

Il est surprenant que la paire de *Chlamydia* montre une augmentation significative des valeurs de Ka avec la distance à l'origine, mais pas du Ks. Il est possible au regard du graphe que quelques gènes à proximité de l'origine de réplication ayant des valeurs de Ks très fortes expliquent cette absence de relation.

Ces résultats confirment ceux obtenus par Sharp *et al*. (Sharp, *et al.*, 1989) chez les enterobactéries et les généralisent à quelques autres espèces de bactéries éloignées phylogénétiquement. Les différentes paires analysées montrent des profils d'augmentation des taux d'évolution assez différents, plutôt linéaire pour *Rickettsia* et *Listeria* ou exponentiel pour les enterobactéries et *Chlamydia*. Il est possible que ces différences témoignent de conséquences de la réplication plus spécifiques à chaque espèce.

3.2.3 ***Discussion***

La répartition phylogénétique des différents profils observés est schématisée fig. 3.13. Étant donnée la représentation phylogénétique du profil d'enrichissement de la région du terminus de réplication, il semble qu'il faille imaginer qu'il a son origine dans un mécanisme très commun, sinon général aux bactéries. La grande représentation des profils plus chaotiques suggèrerait cependant, selon cette hypothèse, qu'il n'existe pas de forte pression de sélection pour imposer cet enrichissement de la région du terminus, et qu'il représente plus probablement la conséquence d'un mécanisme lié à la réplication. Ces profils complexes pourraient dériver des premiers, soit à la suite de réarrangements, soit du fait de transferts ectopiques de larges fragments d'ADN exogène. Dans certains cas, la forme particulière du profil s'explique simplement par l'insertion d'un large plasmide. Le profil en dents de scies de *D. radiodurans* correspond à l'insertion dans le chromosome d'une copie du mégaplasmide riche en A+T également présent dans la cellule, ce qui conduit à deux régions de taux de G+C très différents (White, *et al.*, 1999). De même, chez *M. meliloti*, une grande région beaucoup plus riche en A+T que le reste du génome possède une forte similarité avec un plasmide retrouvé chez plusieurs espèces de protéobactéries proches (Kaneko, *et al.*, 2000), suggérant que ce plasmide s'est inséré récemment dans le chromosome. Pour les autres bactéries cependant, la répartition non aléatoire des gènes en fonction de leur taux de G+C3 semble nécessiter des scénarios plus complexes.

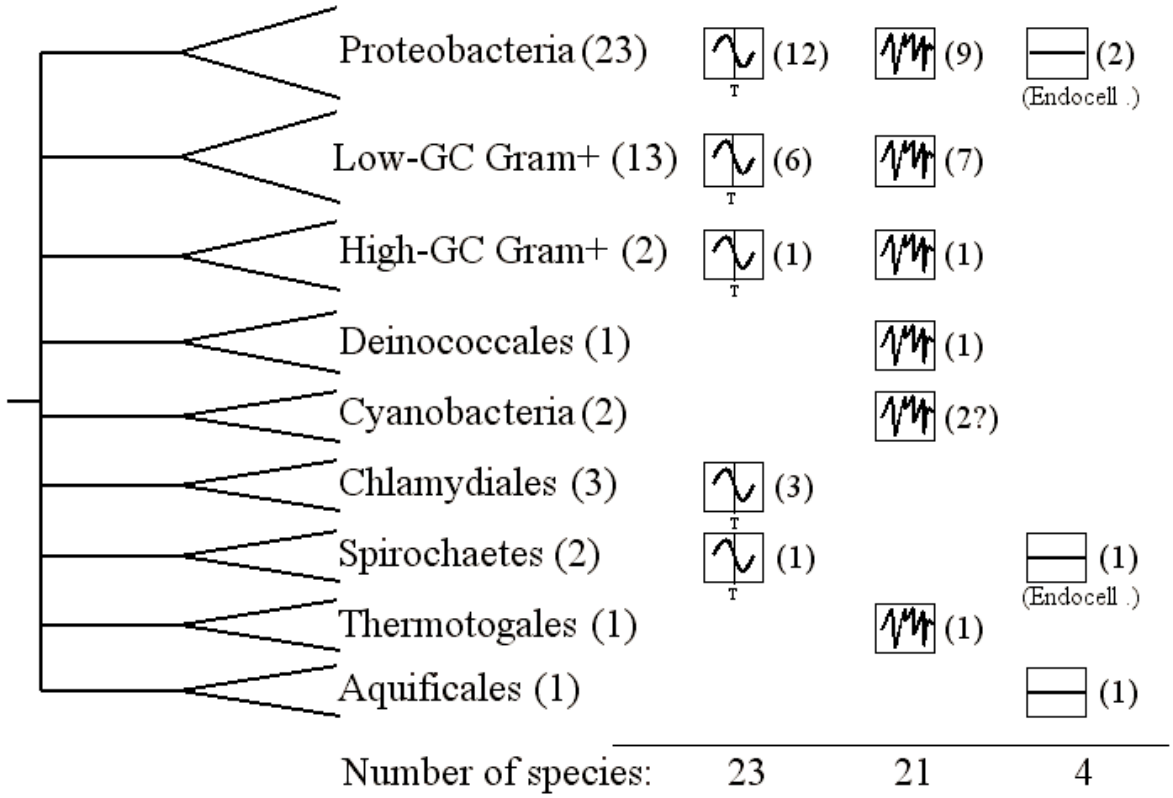


Fig. 3.13 : Représentation phylogénétique des différents profils observés chez les bactéries. Le nombre d'espèces testées est indiqué pour chaque groupe. Les profils sont classés en trois classes : ceux montrant une structuration significative et un enrichissement en A+T de la région du terminus (à gauche) ; ceux montrant une structuration significative mais pas d'enrichissement (centre) ; et ceux ne montrant pas de structuration (droite). Le nombre de génome montrant chaque profil est indiqué dans chaque groupe.

Les deux chromosomes de la bactérie pathogène des plantes *Ralstonia solanacearum* sont tous les deux à la fois très perturbés et très structurés. Ces profils peuvent être mis en relation avec la capacité particulière de cette bactérie à faire de la transformation naturelle (Bertolla, *et al.*, 1997). D'autre part, Brumbley *et al.* (Brumbley, *et al.*, 1993) ont mis en évidence d'importants réarrangements génomiques se produisant spontanément chez cette espèce. Salanoubat *et al.* (2001) ont déjà noté l'existence de grandes régions génomiques contrastant au niveau de leur contenu en G+C et de leur usage du code, et mis cette structure mosaïque du génome en relation avec la grande capacité de *Ralstonia* à intégrer de l'ADN exogène.

Un cas particulièrement intéressant est celui des deux espèces proches *Mycoplasma genitalium* et *M. pneumoniae.* La première montre un profil très intrigant déjà noté par Kerr *et*

al. (Kerr, *et al.*, 1997). Il existe en effet chez cette bactérie une variation régulière du taux de G+C le long du chromosome, mais qui contrairement à la plupart des bactéries montrant un profil similaire ne semble pas être liée à la position de l'origine et du terminus de réplication (Kerr, *et al.*, 1997). De manière surprenante, l'origine aurait plutôt tendance à se trouver dans un région plus riche en A+T (18,4 % de G+C3) que le reste du génome alors que le terminus est dans une région dont le taux de G+C3 est proche de la moyenne du génome (23,3 % de G+C3). La seconde (*M. pneumoniae*) possède un profil beaucoup plus perturbé. Bien qu'elles soient toutes deux des parasites obligatoires, il semble que la dépendance de *M. genitalium* à son hôte soit beaucoup plus forte car il est plus difficile de la cultiver en milieu artificiel. Ceci peut être mis en relation avec le fait que *M. genitalium* possède un génome beaucoup plus réduit (466 gènes) et un taux de G+C beaucoup plus bas que *M. pneumoniae*. Cependant, Himmelreich *et al.* (Himmelreich, *et al.*, 1997) ont montré que l'ordre des gènes était bien conservé en six fragments entre les deux espèces. Les réarrangements ne correspondent pas, contrairement à ce que l'on observe chez la plupart des paires de génomes bactériens, à des inversions symétriques par rapport à l'origine et au terminus de réplication (fig. 3.14). Himmelreich *et al.* (Himmelreich, *et al.*, 1997) ont montré que ces réarrangements étaient liés à la présence de répétitions qui semblent n'avoir été conservées dans leur intégralité que chez *M. pneumoniae* et a proposé que les réarrangements avaient eu lieu dans la lignée de *M. genitalium*. Il semble en effet que de tels réarrangements puissent jouer un rôle très important chez ces bactéries, en relation avec le mécanisme de virulence (Rocha et Blanchard, 2002). Cependant, le profil de *M. genitalium*, quel que soit le mécanisme à son origine, suggère une grande stabilité du génome car des réarrangements perturberaient l'aspect régulier du profil. On peut à cet égard se poser la question de savoir si le profil trouvé chez *M. genitalium* ne pourrait pas être ancestral et avoir été modifié par des réarrangements ayant eu lieu plutôt dans la lignée de *M. pneumoniae*, contrairement à ce qu'ont proposé Himmelreich *et al.* (Himmelreich, *et al.*, 1997). L'hypothèse alternative étant une mise en place rapide du profil chez *M. genitalium* depuis la séparation des deux espèces. L'analyse des différences de G+C3 entre les deux espèces suggère très clairement que ce profil s'est mis en place dans la lignée de *M. genitalium* de manière concomitante avec la forte réduction de son taux de G+C. En effet, la

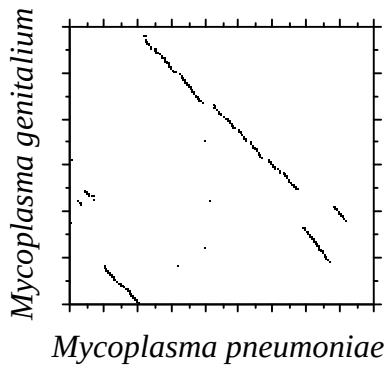
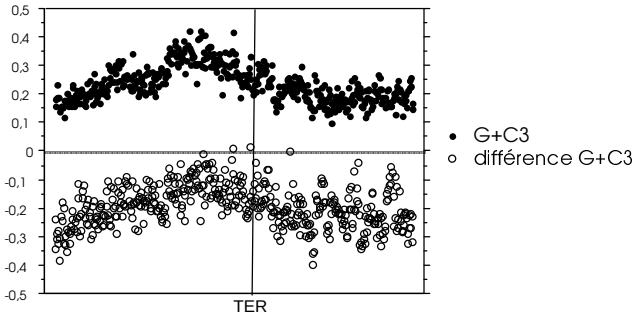


Fig. 3.14. : Dot-plot des génomes de *M. genitalium* et *M ; pneumoniae*. L'ordre des gènes est assez bien conservé mais ne correspond pas aux inversions observées chez la plupart des bactéries



Mycoplasma genitalium

Fig. 3.15. : G+C3 le long du génome de *Mycoplasma genitalium* (points noirs) et différence entre le G+C3 de *M. genitalium* et *M. pneumoniae* ($G+C3_{Mg}$-$G+C3_{Mp}$) (points blanc). La différence est toujours négative ce qui est consistent avec l'enrichissement en A+T de *M. genitalium*. Cependant, la région la plus pauvre en G+C chez *M. genitalium* correspond également à celle présentant les plus fortes différences.

fig. 3.15. montre que les régions les plus riches en G+C du génome de *M. genitalium* sont celles dont le taux de G+C a les plus varié depuis la séparation avec *M. pneumoniae*. Cette structuration particulière, ainsi que le patron de réarrangement du génome suggère que le mécanisme de réplication chez *M. genitalium* possède des propriétés atypiques chez les bactéries.

3.2.3.1 L'hétérogénéité des taux d'évolution : mutation ou sélection différentielle ?

L'augmentation des taux d'évolution dans la région du terminus peut s'expliquer soit par une différence d'efficacité de la sélection, soit par une différence de taux de mutation. Une localisation biaisée des gènes dont l'usage du code est très contraint par la sélection à proximité de l'origine de réplication pourrait en effet donner les résultats de la fig. 3.12. Sharp *et al.* (Sharp, *et al.*, 1989) n'avaient pas trouvé de corrélation entre le niveau d'expression des gènes et leur proximité à l'origine chez les enterobactéries. Il semble cependant que les variations de CAI soient nettement corrélées à la distance à l'origine chez *E. coli* et *S. typhimurium* (fig. 3.9). Mais il semble que la structuration du CAI soit due essentiellement à la variation de G+C3 : chez de nombreuses bactéries comme les entérobactéries, *B. melitensis, V. cholerae, S. meliloti* et *M. tuberculosis* où G+C3 et CAI sont positivement corrélés, les valeurs de CAI sont faibles dans la région du terminus, mais à l'inverse, elle sont fortes chez les bactéries dont les codons optimaux ont tendance à être riches en A+T (*B. subtilis, L. monocytogenes, S. aureus*…). Chez ces dernières bactéries, le biais mutationnel à proximité du terminus et la sélection traductionnelle vont dans le même sens pour le choix des codons.

Une manière de tester si la variation du Ks le long du génome peut être attribuée au regroupement des gènes fortement exprimés à proximité de l'origine de réplication, est de

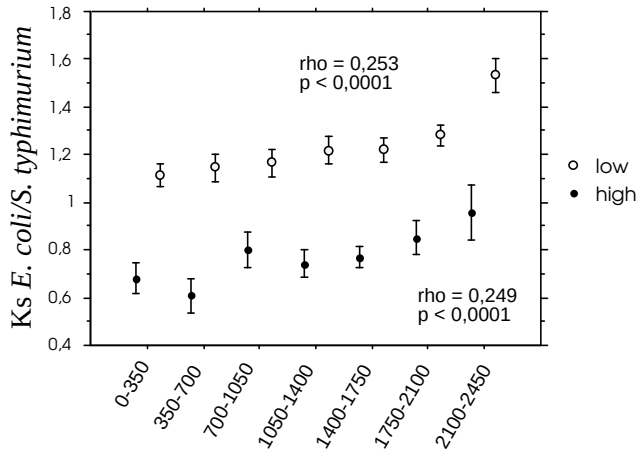


Fig. 3.16 : Taux d'évolution des gènes en fonction de la distance à l'origine pour les gènes ayant un fort CAI (CAI > 0,5 - points noir) et ceux ayant un faible CAI (CAI < 0,5 -points blancs) chez les enterobactéries. La même observation est valable pour différents seuils de CAI. Voir aussi légende de la fig. 3.12.

séparer les gènes présentant de fortes et de faibles valeurs de CAI. Si la classe des gènes ayant un fort CAI peut être considérée comme hétérogène car certains gènes peuvent montrer de fortes valeurs de CAI du fait du biais mutationel, la classe des gènes ayant de faibles valeurs de CAI doit refléter plus fidèlement le patron de mutation le long du génome. L'augmentation du Ks avec la distance à l'origine est fortement significative ($p < 10^{-4}$) pour les deux classes de gènes chez les enterobacteries (fig. 3.16), *Listeria*, et *Rickettsia*. Ceci exclut donc l'hypothèse d'une variation du Ks due au regroupement des gènes fortement exprimés.

Les valeurs particulièrement élevées de Ks dans la région du terminus des espèces montrant un enrichissement en A+T de cette région (entérobactéries, *Listeria*, *Rickettsia* et *Chlamydia*) ainsi que celles du Ka chez les entérobactéries et *Chlamydia* suggèrent que cette région du génome subit, au moins chez ces espèces, des contraintes particulières qui se traduisent par un taux de substitution plus élevé vers les bases A et T.

3.2.3.2 Des contraintes particulières dans la région du terminus ?

Plusieurs hypothèses peuvent expliquer une augmentation du taux de A+T3 à proximité du terminus. Médigue *et al.* (Medigue, *et al.*, 1991) et Lawrence et Ochman (Lawrence et Ochman, 1997; Lawrence et Ochman, 1998) ont noté par exemple que les gènes détectés comme ayant été acquis récemment chez *E. coli* avaient tendance à être sur-représentés dans la région du terminus de réplication et à être plus riches en A et T que le reste du génome. De plus, chez *B. subtilis*, plusieurs prophages sont insérés dans cette région (Kunst, *et al.*, 1997) Ainsi, on peut imaginer que le terminus de réplication constitue un site préférentiel d'insertion des éléments étrangers dans le génome et que ces gènes transférés ont

tendance à être plus riches en A+T que le génome hôte. Rocha et Danchin (Rocha et Danchin, 2002) ont en effet montré que les éléments parasites des génomes comme les phages, les plasmides et les IS présentaient une tendance systématique à être plus riches en A+T que leur génome hôte. Pour tester cette hypothèse, nous avons tracé le profil de *Salmonella* et *Escherichia* en ne prenant en compte que les gènes étant déjà présents chez leur ancêtre commun et ayant conservé leur position (Fig. 3.17). Le même enrichissement en A+T de la région du terminus de réplication est visible et ce, également pour *Chlamydia*, *Listeria* et *Rickesttsia prowazekii*. La différence avec le reste du génome semble même plus marquée. Il faut imaginer dans ce cas que l'insertion des gènes étrangers provoque une augmentation du taux de substitutions vers A+T dans cette région pour les gènes « résidents ». Cependant, l'hypothèse d'un enrichissement en A+T dû aux transferts explique difficilement le cas des *Chlamydia* ou *Mycobacterium tuberculosis* chez qui peu de gènes ayant été acquis récemment par transferts horizontaux ont été détectés (Garcia-Vallve, *et al.*, 2000; Ochman, *et al.*, 2000). Il est également possible que les méthodes de prédiction des gènes transférés horizontalement utilisées par Lawrence et Ochman (Lawrence et Ochman, 1997) ou Médigue *et al.* (Medigue, *et al.*, 1991) surestiment le nombre de gènes récemment acquis dans la région du terminus de réplication chez *E. coli* à cause d'un biais de mutation intrinsèque. Nous reviendrons sur ce point dans la section 3.2.3.3.

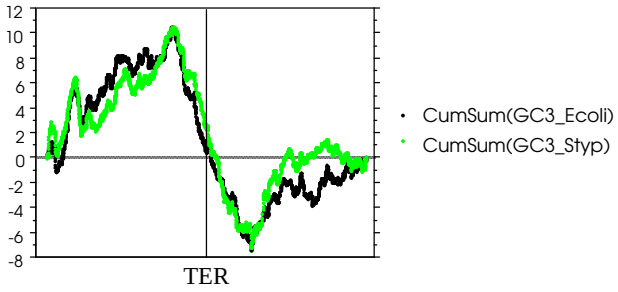


Fig. 3.17 Graphe cumulé des valeurs de G+C3 en n'utilisant que les gènes présents chez l'ancêtre commun à *E. coli* et *S. typhimurium* et ayant conservé leur position dans le génome. Les deux génomes sont représentés (*E. coli* en noir et S. *typhimurium* en clair) et montrent des profils identiques et un fort enrichissement en A+T de la région du terminus.

Une autre hypothèse possible considère les contraintes structurales qui s'exercent sur la région particulière du chromosome qu'est le terminus de réplication. Plusieurs problèmes se posent en effet dans cette région à la fin de la réplication : les deux fourches de réplication doivent se rencontrer au niveau du site *dif*, ce qui nécessite parfois l'arrêt ou le ralentissement d'une des deux fourches au niveau des sites *ter* (Bussiere et Bastia, 1999; Wake, 1997); des contraintes structurales fortes peuvent s'exercer sur cette région du fait de la rencontre des deux fourches (Lewis, 2001) ; les caténats et les dimères de chromosomes doivent y être résolus (Lemon, *et al.*, 2001; Perals, *et al.*, 2001; Lewis, 2001) ; la région du terminus pourrait

jouer un rôle dans la ségrégation des chromosomes néo-synthétisés dans les deux cellules filles notamment en interagissant avec les protéines XerCD et FtsK (Perals, *et al.*, 2001)... Capiaux *et al.* (Capiaux, *et al.*, 2001) ont montré que certains oligomères tendent à augmenter en fréquence à proximité du terminus. Ussery *et al.* (Ussery, *et al.*, 2001) ont également montré que FIS (une protéine architecturale très abondante chez *E. coli*, notamment pendant les phases exponentielles de croissance, et composant essentiel de la chromatine - Finkel et Johnson, 1992; Schneider, *et al.*, 2001; Travers, *et al.*, 2001) possède une forte densité de sites dans une région d'approximativement un Megabase autour du terminus d'*E. coli*. Cette région correspond presque exactement à la région d'enrichissement en A+T. Il a également été montré que cette même région est enrichie en séquences favorisant la courbure de l'ADN chez *E. coli* et *B. subtilis* (Pedersen, *et al.*, 2000), ce qui pourrait jouer un rôle dans la fixation d'autres proteines comme H-NS, elle aussi impliquée dans la condensation de l'ADN (Ussery, *et al.*, 2001). La structure particulière de cette grande région entourant le terminus pourrait jouer un rôle dans la ségrégation des chromosomes (Tsai et Sun, 2001) et/ou la résolution des dimères de chromosomes au niveau du site *dif*, dont il a été montré que le rôle ne pouvait être assuré qu'en présence d'une large part de ses séquences flanquantes (Perals, *et al.*, 2000). Il est généralement argumenté que le mécanisme de la terminaison de la réplication d'*E. coli* et *B. subtilis* sont apparus indépendamment (Hill, 1992; Wake, 1997) bien qu'ils soient basés sur des mécanismes extrêmement similaires (Wake, 1997; Bussiere et Bastia, 1999). Bien qu'aucun homologue de FIS ne semble avoir été trouvé chez *B. subtilis*, une protéine appelée AbrB lui ressemble beaucoup en terme de taille, de fixation à l'ADN, de patron d'expression, et du contrôle qu'elle joue sur l'expression des autres gènes. Ceci suggère qu'elle pourrait jouer le même rôle que FIS chez cet organisme (O' Reilly et Devine, 1997). Ainsi, la richesse en A+T de la région du terminus pourrait avoir un intérêt fonctionnel pour le processus de réplication, en facilitant la fixation de protéines et la formation de boucles au moins chez *E. coli* et *B. subtilis*. Il se pourrait dans ce cas que l'augmentation des taux d'évolution dans cette région soient le témoignage d'un conflit entre deux niveaux de sélection : celui du gène (dont la composition en codon et en acides aminés est contrainte) et celui du chromosome (dont les charactèristiques structurales pourraient être soumises à sélection dans cette région). Bien que le rôle des sites *ter* dont la fonction est d'empêcher les fourches de réplication de dépasser le site de la terminaison soit bien connu et que ces sites apparaissent chez de nombreuses espèces, il est intéressant de noter que leur délétion n'a aucun effet détectable sur la fitness ni d'*E. coli* ni de *B. subtilis* au moins dans les conditions de laboratoire (Bierne et Michel,

1994). Ceci suggère que d'autres espèces peuvent avoir développé des systèmes alternatifs pour la terminaison de leur réplication.

Une autre possibilité est que ces contraintes s'exerçant à proximité du terminus aient des effets mutagènes. Par exemple, une fourche de réplication arrêtée est caractérisée par la présence notamment de région d'ADN simple brin persistantes. Il est donc possible qu'à proximité des site *ter*, les séquences soient plus sensibles aux processus de mutation et de recombinaison (Bierne, *et al.*, 1997). D'autre part, la région du terminus de réplication peut également posséder des mécanismes de réparation de l'ADN différents du reste du chromosome. C'est l'hypothèse que privilégient Sharp *et al.* (Sharp, *et al.*, 1989) pour expliquer la corrélation entre la distance au terminus et l'augmentation des taux d'évolution chez les entérobactéries. Ils proposent que la présence de fourches multiples à proximité de l'origine de réplication pendant le phase exponentielle de croissance permet de réparer les lésions par recombinaison plus fréquemment que dans la région du terminus. En effet, les séquences proches du terminus doivent selon ce modèle se trouver moins souvent en plusieurs copies dans la cellule et donc avoir moins d'opportunités d'être réparées en utilisant la recombinaison. Cependant, ce modèle correspond mal à celui de la « replication factory » décrit section 3.1.1, où les séquences proches de l'origine sont, immédiatement après leur réplication, attirées vers les pôles opposés de la cellule. Le mécanisme de réparation par recombinaison implique plus probablement les deux brins qui viennent d'être synthétisés, et il n'existe pas de ce point de vue de différence entre une fourche proche de l'origine et proche du terminus. Cependant, les mécanismes de réparation des lésions pendant la réplication peuvent tout de même présenter des différences entre la région de l'origine et celle du terminus. Lorsqu'une lésion de l'ADN est rencontrée par le complexe protéique assurant la réplication, celui-ci est arrêté. Pour que la réplication soit ré-initiée, la lésion doit être soit réparée, soit passée par le complexe. L'un et l'autre de ces mécanismes nécessitent la régression de la fourche de réplication, c'est-à-dire le désappariement des brins néosynthétisés de leurs matrices, le ré-appariement des brins matrices entre eux et l'appariement des deux brins néosynthétisés (voir Fig. 3.18). Cette étape nécessite l'activité d'hélicases (RecG et PriA) notamment pour désapparier les brins neosynthétisés de leurs matrices (Gregg, *et al.*, 2002; McGlynn et Lloyd, 2002). Or, cette activité hélicase se produit dans le sens inverse de la réplication. Cela ne pose pas de problème particulier à proximité de l'origine, mais dans le piège que constituent les sites *ter*, les hélicases peuvent être empêchées de procéder du fait de

l'action polaire des protéines Tus. Il est alors possible que le seul moyen pour la réplication de continuer sa progression soit d'introduire une mutation en face de la lésion (translésion).

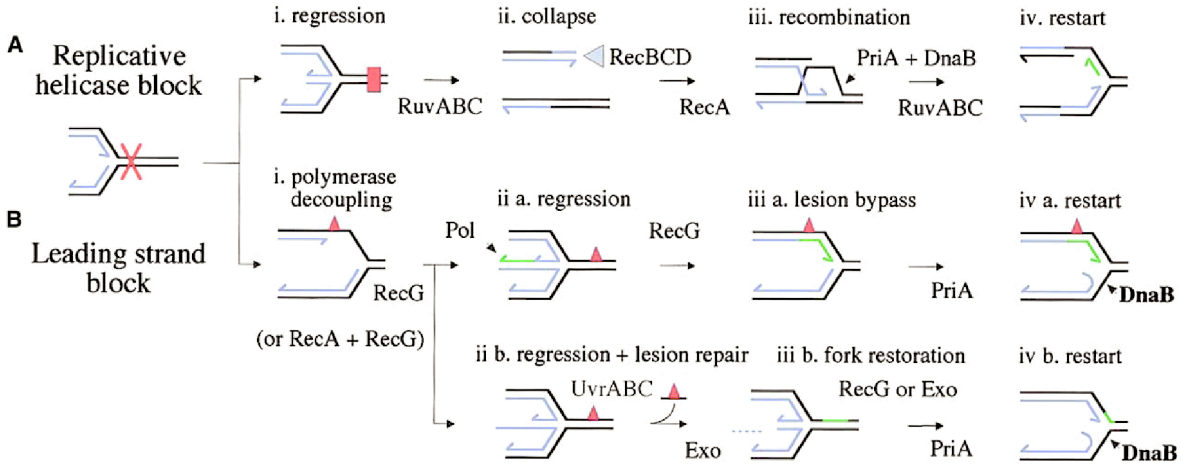


Fig. 3.18 : Modèle de « sauvetage » d'une fourche de réplication bloquée face à une lésion de l'ADN. Une lésion double brin nécessite un réparation par recombinaison (A) alors qu'une lésion simple brin peut être réparé (Bb) ou passée (Ba) par régression de la fourche (Bii). Dans tous les cas, les doubles brins qui viennent d'être synthétisés sont désappariés pour former un intermédiaire de type jonction de Holliday (Ai et Bii) ce qui nécessite l'action d'une hélicase dans la direction opposée à la réplication. Une fois la lésion passée, la réplication doit être ré-initiée. Le rôle des hélicases RecG et PriA n'est pas encore complètement élucidé. Extrait de Gregg, *et al.*, 2002.

Ce mécanisme qui serait alors utilisé préférentiellement dans la région du terminus est à la fois générateur d'erreurs et biaisé vers A+T. Les polymérases impliquées dans la translésion (notamment polII et polV) qui, face à une lésion de l'ADN et en absence de recombinaison, permettent à la synthèse d'ADN de se poursuivre en introduisant des erreurs, respectent en effet la règle d'incorporation préférentielle d'un dAMP en face d'un site abasique (la « A-rule ») (Strauss, 1991; Ide, *et al.*, 1995). Ce type de lésion se produit fréquemment dans les cellules, aussi bien spontanément qu'en présence d'agents mutagènes, ou bien par l'action de certaines N-glycosylases après reconnaissance d'une base modifiée (Ide, *et al.*, 1995). L'incorporation préférentielle de A aux sites abasiques pourrait donc expliquer l'enrichissement d'une région où la réparation impliquant un régression de la fourche de réplication est rare.

Ce mécanisme semble être en contradiction avec un travail récent mené par Hudson *et al.* (Hudson, *et al.*, 2002). Ces auteurs ont inséré un gène *LacZ* non fonctionnel à différents

locus du génome de *S. enterica* et ont mesuré les différences dans les fréquences de réversion vers l'allèle fonctionnel. Ils ne trouvent pas de différence dans les taux de réversion entre des gènes insérés à proximité de l'origine et du terminus. Plusieurs types de réversions ont été analysées et en particulier, il se sont intéressés aux différences entre les transitions et les tranversions. Cependant, l'ensemble des réversions analysées consiste en des mutations de A ou T vers G ou C. Ainsi, si le taux de mutation vers A+T augmente à proximité du terminus, leur étude ne permet pas de le détecter.

Une possibilité pour montrer la tendance du terminus à s'enrichir vers A+T serait d'utiliser trois génomes complets de souches relativement proches de la même espèce. Par une étude de parcimonie, il serait en effet possible d'orienter les substitutions chez deux des trois espèces et ainsi de voir si la fréquence des différents types de substitution varie le long du génome. Malheureusement, un cas suffisamment favorable n'existe pas encore : par exemple, les deux souches *E. coli* O157:H7 sont trop proches pour que les gènes aient accumulé assez de différences. D'autre part, si la comparaison de *E. coli* O157:H7 et *E. coli* K12 montre des différences suffisamment importantes, le racinement de ce groupe n'est possible pour l'instant qu'au moyen d'une des deux souches de *Salmonella*, dont les valeurs de Ks sont souvent très supérieures à 1, ce qui exclut de pouvoir utiliser l'hypothèse de parcimonie.

3.2.3.3 L'implication pour les méthodes de détection des transferts horizontaux.

Comme l'a montré Ragan (Ragan, 2001), les différentes méthodes de détection des transferts horizontaux donnent des résultats très différents chez *E. coli* et prédisent même parfois des ensembles de gènes non recouvrant. Les résultats que nous présentons ici montrent que l'une des hypothèses fortes des méthodes intrinsèques, celle concernant la faible hétérogénéité du contenu en bases des gènes (Lawrence et Ochman, 1997) d'un génome procaryote n'est pas respectée pour la plupart des espèces. La structuration du G+C3 et du CAI montrent que différentes parties du génome peuvent avoir des usages du code différents. Ceci implique qu'un biais est possible dans les méthodes de prédiction des gènes transmis horizontalement basée sur la composition en codons, qui sont les plus utilisées (Medigue, *et al.*, 1991; Lawrence et Ochman, 1997; Lawrence et Ochman, 1998; Moszer, *et al.*, 1999; Garcia-Vallve, *et al.*, 2000; Ochman, 2001). En effet, si une telle structuration existe, il

devient par exemple complètement injustifié de considérer *a priori* que les taux de G+C des gènes natifs d'un génome suivent une distribution normale (voir section 1.6). Comme nous l'avons déjà noté, Lawrence et Ochman (Lawrence et Ochman, 1997; Lawrence et Ochman, 1998) ont remarqué que les gènes qu'ils détectent comme étant d'origine étrangère chez *E. coli* sont significativement plus représentés dans la région du terminus. Le profil observé chez *E. coli* semble pouvoir être représenté dans tous les phylums bactériens ce qui suggère que ce biais peut avoir un impact fort sur toutes les estimations faites à ce jour des pourcentages de gènes étrangers dans les génomes.

3.3 Étude de l'usage du code des gènes transférés horizontalement

Du fait de l'existence d'un biais mutationnel et d'un usage du code spécifique à chaque espèce, les génomes bactériens sont considérés comme étant homogènes au niveau de la composition de leurs gènes (Lawrence et Ochman, 1997; Lawrence et Ochman, 1998). Les seuls facteurs considérés sont la plupart du temps le taux d'expression des gènes et dans certains cas extrêmes comme *Borrelia burgdorferi*, le brin codant (direct ou retardé) et la composition en acides aminés. Ces facteurs pris en compte, les gènes montrant une composition (en bases, en codons ou en oligonucléotides) atypique sont considérés comme étant issus de transferts horizontaux. Leur inadéquation au reste du génome s'explique par le fait qu'ils portent encore la trace des biais mutationnels et de l'usage du code de leur précédent hôte. Il en résulte, selon Lawrence et Ochman (Lawrence et Ochman, 1997), que ces gènes présentent de fortes valeurs de χ^2, c'est-à-dire que leur usage du code est biaisé, mais de faibles valeurs de CAI, c'est-à-dire que le biais ne va pas dans le sens des gènes fortement exprimés du nouvel hôte. Lawrence et Ochman (Lawrence et Ochman, 1997) proposent que de tels gènes subissent une « amélioration », c'est-à-dire qu'il s'adaptent aux biais mutationnel et d'usage du code de leur nouvel environnement.

Dans leur analyse multivariée de l'usage du code des gènes d'*E. coli*, Médigue *et al.* (Medigue, *et al.*, 1991) montrent que le génome peut être séparé en trois classes de gènes cohérentes au niveau de leur usage du code. En effet, ils utilisent une méthode de classification dite des centres mobiles permettant de regrouper un ensemble de points en un nombre arbitrairement choisis de classes cohérentes. L'une des classes est constituée des gènes fortement exprimés (usage du code biaisé par l'abondance des ARNt), une autre

représente les gènes faiblement exprimés (usage du code biaisé par le biais mutationnel), et une troisième est constituée de gènes de fonctions inconnues et d'éléments étrangers comme les séquences d'insertion (IS), ou des phages. Cette troisième classe est donc interprétée comme la classe des gènes inadaptés au génome d'*E. coli* et ayant été acquis récemment par transferts horizontaux. Cette interprétation fait une hypothèse qui n'est pas mentionnée explicitement par Médigue *et al.* (Medigue, *et al.*, 1991) : les gènes acquis récemment, constituent un ensemble homogène au niveau de leur usage du code, en comparaison du reste du génome. Ceci est surprenant si l'on considère que ces gènes, qui représentent plus de 10 % du génome, proviennent d'un certain nombre d'événements de transferts indépendants, impliquant différentes espèces. Médigue *et al.* notent en effet que ces gènes ont une forte tendance à être plus riches en A+T que le reste du génome. Le premier axe de l'analyse est en effet très fortement corrélé au G+C3 des gènes (Rho = 0,666 ; $p < 10^{-4}$) (Fig. 3.19) mais également, bien que moins fortement, à la distance à l'origine (Rho = -0,191 ; $p < 10^{-4}$). Ce résultat pose la question de savoir ce qui confère à ces gènes des caractéristiques communes : les gènes transférés proviennent-ils d'un nombre limité de génomes donneurs ? Y a t'il un biais d'incorporation des séquences riches en A+T chez *E. coli* ? Ou encore ces gènes portent-ils la trace d'autre chose que leur hôte précédent ?

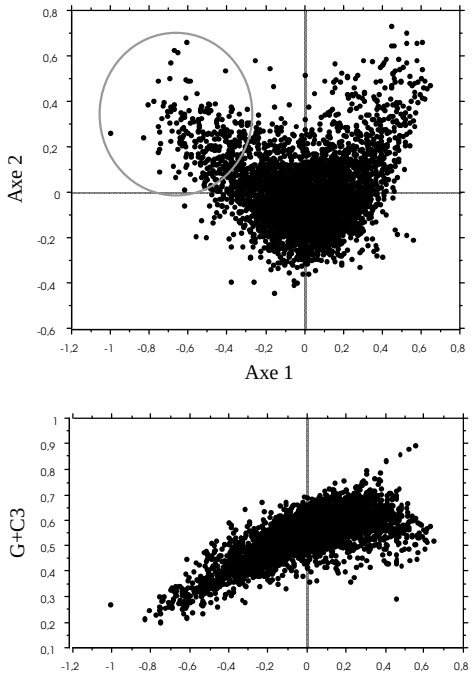


Fig. 3.19 : AFC sur les fréquences relatives des codons de l'ensemble des gènes du génome d'*E. coli* (en haut) et relation entre le premier axe de l'analyse et le taux de G+C en troisième position des codons. Les gènes consituant la classe des gènes transférés (cercle) sont principalement séparés des autres par le premier axe. Ils correspondent à des gènes riches en A+T dont la distribution en G+C3 est assez resserrée.

3.3.1 ***Matériels et Méthodes***

3.3.1.1 Principe de la détection des gènes acquis et perdus récemment

Pour tenter d'éclaircir ce point, nous avons décidé d'étudier l'usage du code des gènes acquis récemment, détectés par une méthode indépendante en collaboration avec Emmanuelle Lerat. La disponibilité de génomes complets proches permet en effet de trouver des gènes présents chez une espèce (ou souche) bactérienne et absents de ses espèces voisines. On peut ensuite utiliser l'hypothèse de parcimonie pour discriminer entre une acquisition récente et des pertes chez les espèces voisines. Le cas idéal permettant d'inférer les événements de gain et de perte est celui présenté fig. 3.20. Les gènes présents chez A n'ayant aucun homologue ni chez B, ni chez C, peuvent soit avoir été présents chez l'ancêtre commun a A, B et C, puis perdus indépendamment chez B et C, soit avoir été acquis chez l'ancêtre commun à A et B puis perdu chez B, soit avoir été acquis récemment chez A. La dernière interprétation paraît plus parcimonieuse si l'on considère que l'acquisition et la perte ont des probabilités comparables. Les modèles expliquant le maintien de la taille des génomes reposent sur cette hypothèse (Mira, *et al.*, 2001) : les acquisitions et les pertes se compensent chez les bactéries libres, et la réduction des génomes parasites s'explique par une diminution des transferts dans les milieux confinés que sont les hôtes. Une manière de vérifier si, dans le cas étudié, cette hypothèse est réaliste est d'estimer également la quantité de gènes perdus. De même, une « perte » apparente peut correspondre à plusieurs interprétations : la perte effective chez l'espèce considérée ou deux acquisitions indépendantes chez les deux autres espèces. Ces estimations peuvent se faire chez A et B. Si le nombre de gènes « perdus » chez B est du même ordre de grandeur ou inférieur au nombre de gènes « acquis » chez A, il est peu probable, même si C a des taux de délétion relativement forts, que les gènes apparemment acquis chez A correspondent à des pertes chez B et C.

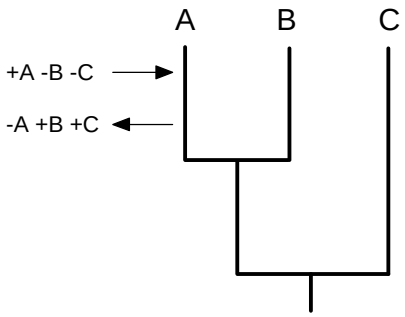


Fig. 3.20 :Description de la méthode de détection des gènes acquis et perdus récemment par BLAST lorsqu'on dispose de trois génomes complets proches. Voir détails dans le texte.

Un autre phénomène que nous n'avons pas encore mentionné peut affecter ces estimations du point de vue quantitatif : les duplications. En effet, si un gène acquis récemment est dupliqué après son transfert, les nouvelles copies grossiront le nombre des gènes prédits comme transférés. Ceci n'a pas beaucoup d'importance dans notre cas car nous ne tentons pas d'estimer le nombre d'évènements de transferts, mais seulement les gènes se trouvant dans un contexte génomique nouveau pour eux. De même, si un gène est perdu dans une lignée et dupliqué dans sa lignée sœur, le nombre de pertes pour ce gène sera surestimé. Encore une fois, ceci ne constitue un biais que du point de vue de la quantification des pertes.

3.3.1.2 Génomes utilisés

Pour conduire l'analyse telle que nous venons de la décrire, nous devons disposer d'au moins trois génomes complètement séquencés et suffisamment proches. Un certain nombre de tels cas sont disponibles chez les bactéries :

- cinq génomes proches sont disponibles dans le groupe des enterobactéries : trois souches d'*E. coli* (K12, O157:H7 EDL933 et O157:H7 Sakai) et deux espèces (ou sous espèces) de *Salmonella* (*S. typhimurium* LT2 et *S. enterica*).

- trois génomes proches sont disponibles dans le groupe des ε-protéobactéries : deux souches d'*Helicobacter pylori* (J99 et 26695) et une souche de *Campylobacter jejunii.*

- trois génomes proches sont disponibles dans le genre *Streptococcus* : deux souches de *S. pneumoniae* (R6 et TIGR4) et une souche de *S. pyogenes.*

D'autres cas, comme par exemple celui des Mycobactéries, n'ont pas été considérés du fait du processus de réduction qui s'exerce sur le génome de *M. leprae* : en effet, la probabilité de pertes étant très élevée chez cette bactérie (elle possède presque deux fois moins de gène que *M. tuberculosis*), l'hypothèse de parcimonie peut difficilement être faite. De même, le cas des Chlamydiales n'a pas été analysé car leur génome subit également une réduction importante.

Toutes les séquences utilisées ainsi que leurs annotations ont été extraites de la banque de génomes complets EMGLib (Perriere, *et al.*, 2000a) en utilisant le système de requête ACNUC (Gouy, *et al.*, 1985). Seules les séquences de plus de 150 pb ont été utilisées. Les gènes de phages et d'IS sont identifiés sur la base de leurs annotations.

3.3.1.3 Détection des gènes récemment acquis

Afin de détecter les gènes ayant été acquis récemment dans l'espèce A (cas +A-B-C dans la fig. 3.20), une requête de BLASTP (Altschul, *et al.*, 1997) des protéines contenant plus de 50 acides aminés du génome A est lancée sur une banque constituée des protéines de B et C. Un gène est considéré comme absent des deux autres génomes si aucune séquence ayant un score (E-value) inférieur à 0,001 n'est détecté dans cette banque. Ce critère de sélection est assez stringent de manière à ne pas prendre en compte de gènes ayant très fortement divergé. De ce fait, le nombre de gènes acquis récemment est sous estimé. Cependant, d'autres seuils de score donnent des nombres de gènes très comparables. Une requête de BLASTN (Altschul, *et al.*, 1997) des séquences nucléiques correspondant à ces gènes sur les génomes complets de B et C est ensuite effectuée pour vérifier que ceux-ci ne sont pas absents du fait d'erreurs d'annotations. Les gènes non détectés comme acquis récemment par cette méthodes sont considérés comme natifs.

3.3.1.4 Détection des gènes perdus

Afin de détecter les gènes ayant été perdus récemment dans l'espèce A (cas –A+B+C dans la figure 3.20), une requête de BLASTP (Altschul, *et al.*, 1997) des protéines contenant plus de 50 acides aminés du génome B est lancée sur une banque constituée des protéines de A et C. Un gène est considéré comme ayant été perdu récemment chez A s'il est absent du génome A ($E > 0{,}001$) et présent dans le génome C ($E < 10^{-20}$). De même, les critères utilisés ici sont très stringents, de manière à ne prendre en compte que des gènes effectivement perdus et sous estiment le nombre de pertes. Mais de même que dans le cas précédent, les variations de ces seuils n'ont qu'un faible impact sur la quantité de gènes identifiés. Une requête de BLASTN (Altschul, *et al.*, 1997) des séquences nucléiques correspondant à ces gènes sur le

génome complet de A est ensuite effectuée pour vérifier que ces gènes ne sont pas absents du fait d'erreurs de prédictions.

3.3.1.5 Analyse de l'usage du code des gènes natifs et transférés.

Les effectifs de chacun des 59 codons synonymes ont été calculés pour chacun des gènes. Le résultat en est une matrice contenant 59 colonnes et autant de lignes que de gènes analysés. Cette matrice peut être utilisée pour faire une AFC (Analyse Factorielle des Correspondances) (Benzécri, 1973) à l'aide du logiciel ADE-4 (Thioulouse, *et al.*, 1997). Il s'agit d'une analyse multivariée souvent utilisée pour étudier l'usage du code (Grantham, *et al.*, 1981; Shields et Sharp, 1989; Medigue, *et al.*, 1991). Elle permet de calculer la position des séquences dans un espace multidimentionnel en fonction de l'usage du code et d'en donner une représentation graphique dans les dimensions qui maximisent leur dispersion. Les gènes ayant un usage du code semblable sont ainsi regroupés. L'analyse se faisant de manière symétrique, il est possible de représenter les codons dans ces mêmes dimensions, ce qui permet de visualiser ceux qui sont responsable des différents regroupements de gènes. Plusieurs classes de gènes sont considérées : gènes natifs, gènes transférés, gènes de phages et gènes de séquences d'insertion. Pour ne pas biaiser l'analyse en sur-représentant une classe par rapport aux autres, et pour faciliter l'interprétation des graphes, un tirage au sort est fait parmi les gènes des classes les plus nombreuses (et notamment la classe des gènes natifs). Différents tirages au sort donnent des résultats très semblables.

3.3.2 ***Résultats***

3.3.2.1 Gènes récemment acquis ou perdus

Les critères utilisés ici pour identifier les gènes transférés ou perdus récemment sont très stringents de manière à minimiser le nombre de faux positifs. De ce fait, comme toutes les méthodes de détections de transferts, cette méthode ne détecte qu'un sous ensemble des gènes transférés horizontalement. Les nombres de gènes présentés sur la fig. 3.21 ne représentent donc que les gènes dont aucun homologue, même lointain n'apparaît dans les génomes

considérés. Comme nous l'avons déjà précisé, ces estimations d'effectifs ne correspondent pas nécessairement à autant d'évènements de transfert du fait de la possibilité pour les gènes d'être dupliqués après leur intégration dans un génome. Cependant, ces chiffres sont intéressants pour la dynamique des génomes concernés. Dans la plupart des cas, le nombre de gènes récemment acquis est très supérieur au nombres de gènes perdus. Or, on considère la plupart du temps que la taille des génomes est relativement stable, et que les acquisitions sont compensées par des pertes (Mira, *et al.*, 2001). Ces différences peuvent tenir au fait que la séquence constitue un « instantané » du génome de l'espèce, et que si les gènes acquis récemment ont tendance à être perdus rapidement après leur acquisition, ces pertes ne peuvent pas être détectées. La différence du nombre de gains et de pertes de gènes peut donc être interprétée comme l'indice d'un « turn over » particulièrement important de ces gènes. On peut cependant supposer par exemple que les souches pathogènes d'*E. coli* O157:H7 (EDL933 et Sakaï) sont dans une phase d'expansion de leur génome car la quantité de gènes ayant été acquis récemment est très importante alors qu'elles ne se sont séparées que très récemment (la séquence de la plupart de leurs gènes est quasiment conservée à 100 % au niveau nucléïque). Ceci est cohérent avec la grande différence de contenu en gènes qui existe entre ces souches et celle d'*E. coli* K12 (de l'ordre de 1000 gènes). La souche *E. coli* K12 présente nettement moins d'acquisitions de gènes de ce type. La forte proportion de gènes perdus dans cette lignée est probablement surestimée du fait de duplications dans la lignée des *E. coli* pathogènes. En effet, les 273 gènes (ou 283 selon la souche d'*E. coli* prise comme groupe frère) inférés comme perdus correspondent en réalité à 173 (ou 159) familles dans HOBACGEN (Perriere, *et al.*, 2000b). De ce fait, dans tous les cas considérés (sauf peut-être celui d'*E. coli* K12), le rapport des acquisitions et des pertes montre que la probabilité d'acquisition d'un gène est au moins égale à celle d'une perte, ce qui permet d'interpréter les chiffres d'acquisition comme des transferts avec un bonne confiance. Dans les cas où le nombre de perte est très faible, on ne peut complètement exclure la possibilité de deux événements d'acquisition indépendants.

Le cas des enterobactéries permet d'identifier deux classes de gènes transférés en fonction de leur date relative d'acquisition : les gènes acquis après la séparation des *E. coli* pathogènes (O157:H7) de *E. coli* K12 et avant la séparation des deux souches d'*E. coli* O157:H7 (que nous appelerons « transferts anciens »), et ceux acquis depuis cette séparation (« transferts récents »).

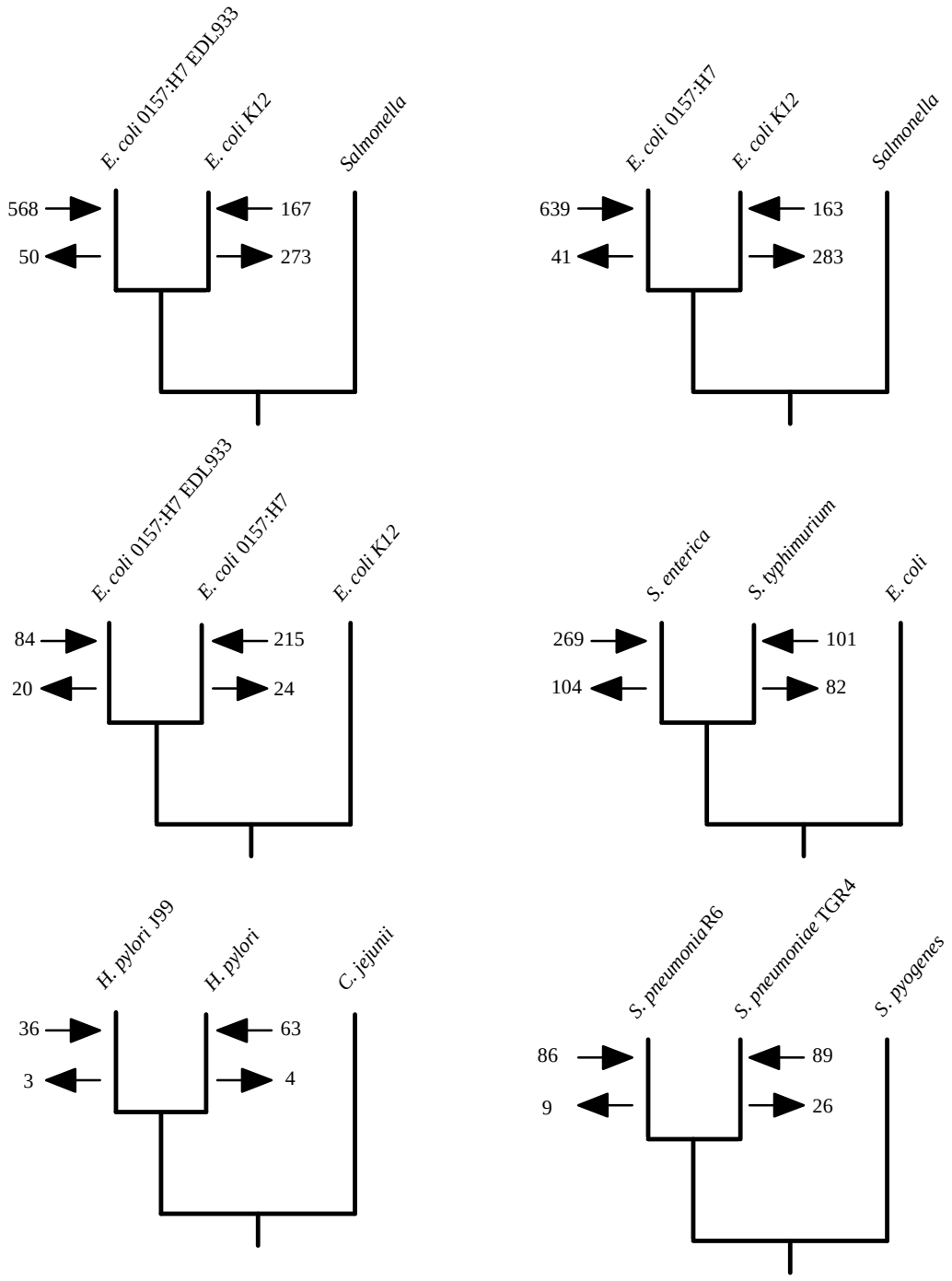


Fig. 3.21 : nombres de gènes acquis (flèche vers la branche) et perdus (flèches vers l'exterieur) dans les différentes lignées étudiées. Pour les enterobactéries, il est possible d'identifier des gènes acquis plus ou moins récemment en comparant les différents arbres de transferts (gènes acquis entre la séparation de K12 des deux O157:H7 et gènes acquis depuis la séparation des deux O157:H7).

Une fraction très importante des gènes détectés par cette méthode ont des fonctions inconnues. Quelques-uns sont annotés comme des protéines membranaires, phages ou IS. Dans les parties suivantes, ces deux dernières classes de gènes apparaîtrons dans les catégories phages et IS.

3.3.2.2 La répartition des gènes récemment acquis

La répartition des gènes récemment acquis dans les génomes peut ici être visualisée indépendamment d'éventuels biais de composition. Dans le chapitre précédent, nous avons en effet montré qu'un biais de composition à proximité du terminus de réplication pouvait potentiellement entraîner des surestimations des gènes transférés dans cette région. En utilisant le test de χ^2 sous l'hypothèse d'équi-répartition des gènes acquis récemment en deux classes (« proches » et « éloignés » du terminus), nous avons trouvé que les deux souches d'*E. coli* O157:H7 présentaient une sur-représentation significative ($p < 0,0025$) des gènes acquis récemment à proximité du terminus dés que la taille de la région considérée comme « proche » est supérieure à 500 kb. Cependant, ceci n'est visible que pour les gènes acquis depuis la séparation d'avec *E. coli* K12, mais pas pour les gènes acquis après la séparation plus récente des deux souches d'*E. coli* O157:H7. Chez *Salmonella enterica*, il faut considérer une région de plus de 600 kb autour du terminus pour voir une sur-représentation significative ($p < 0,0001$). De manière intéressante, dans les génomes de *Salmonella enterica* et d'*E. coli* O157:H7, les gènes prédits comme ayant été perdus chez les espèces voisines sont également sur-représentés dans la région du terminus de réplication. Dans les autres génomes, et notamment chez *E. coli* K12, les gènes acquis ou perdus récemment ne semblent pas être significativement regroupés dans le région du terminus. Il est possible que ce résultat soit dû au fait que les nombres inférés de gènes transférés horizontalement sont relativement faibles pour ces génomes. Cependant, chez *E. coli* O157 :H7 et *S. enterica*, les gènes acquis récemment ne sont pas représentés de manière

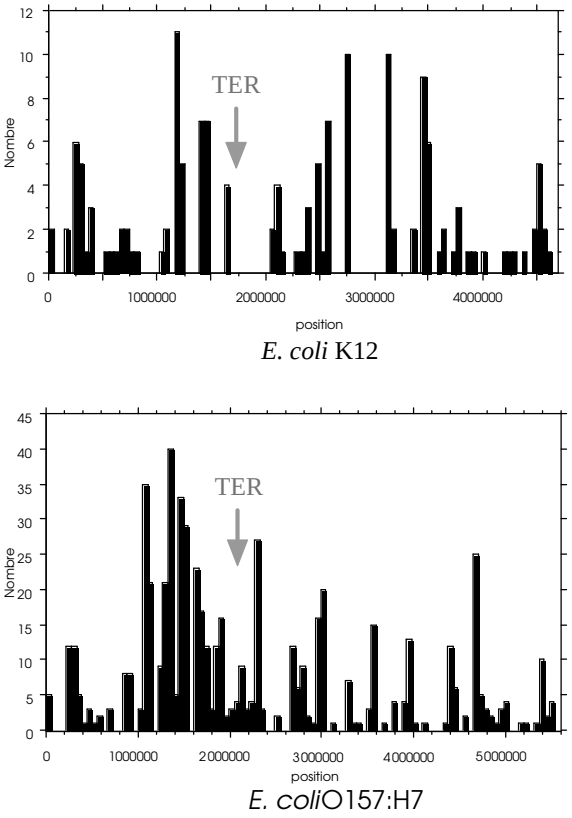


Fig. 3.22 :répartition des gènes récemment acquis détectés par notre méthode dans les génomes de E. coli K12 et O157:H7 EDL933 depuis leur séparation. Seule O157:H7 EDL933 montre une sur-représentation de ces gènes dans la région du terminus

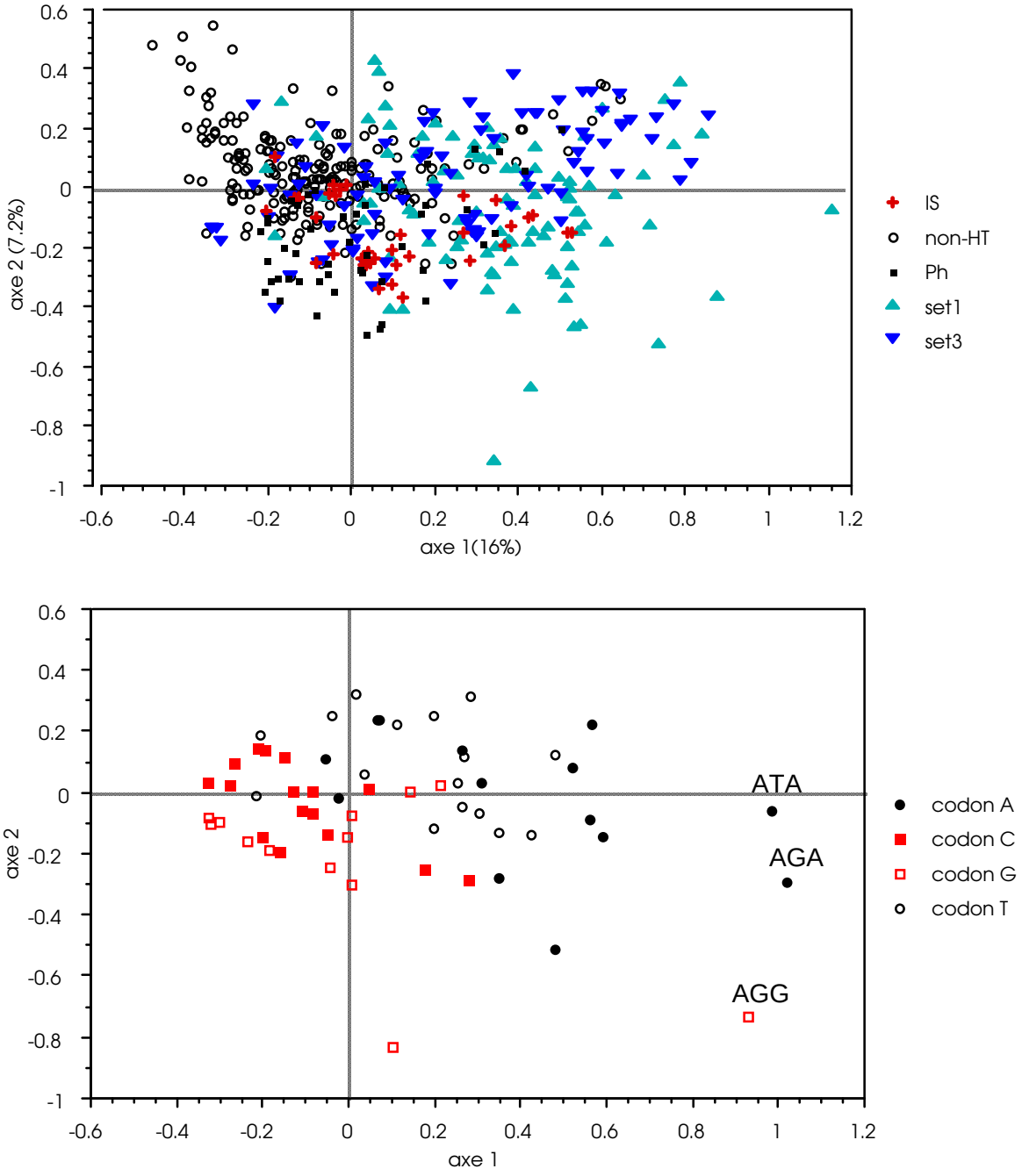
symétrique par rapport au terminus (fig. 3.22), ce qui suggère que ce qui est observé n'est pas réellement lié à la présence du terminus dans cette zone.

3.3.2.3 Analyse du code des gènes transférés horizontalement par l'AFC

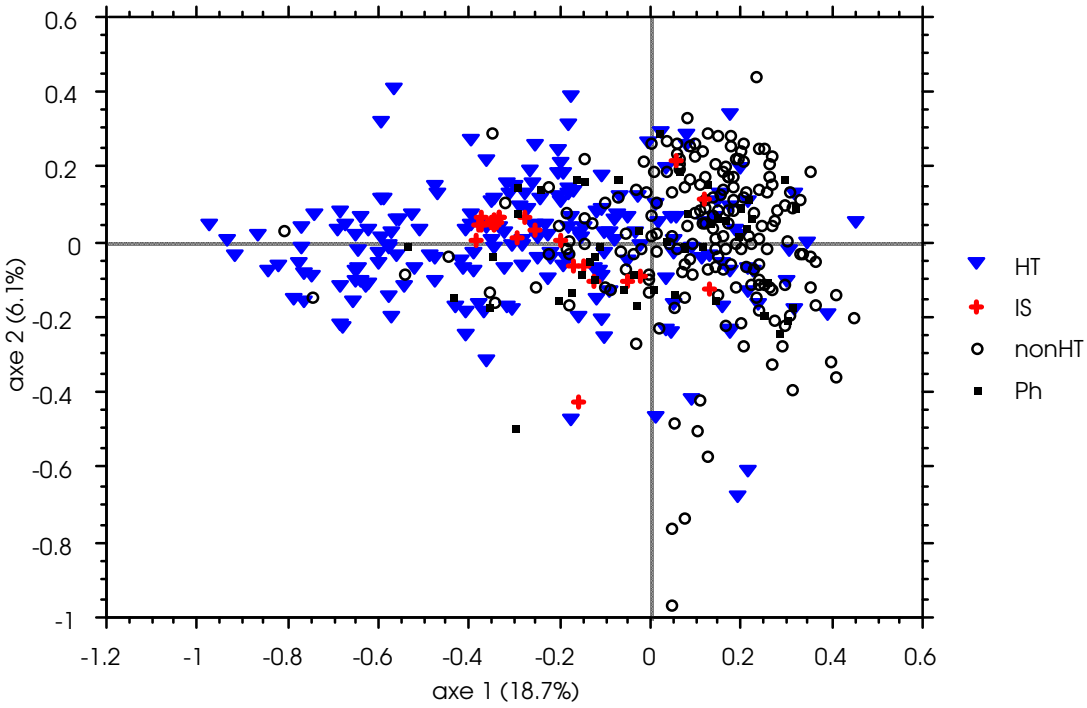
La Fig. 3.23 montre les AFC pour les quatre genres analysés. La projection des gènes ainsi que celle des codons est présentée. Nous avons, en plus des gènes natifs et des gènes transférés horizontalement, ajouté, lorsqu'il en existait dans le génome, des gènes d'éléments parasites du génome comme les phages ou les IS. Dans toutes ces analyses, le premier axe sépare bien les gènes récemment acquis des gènes natifs.

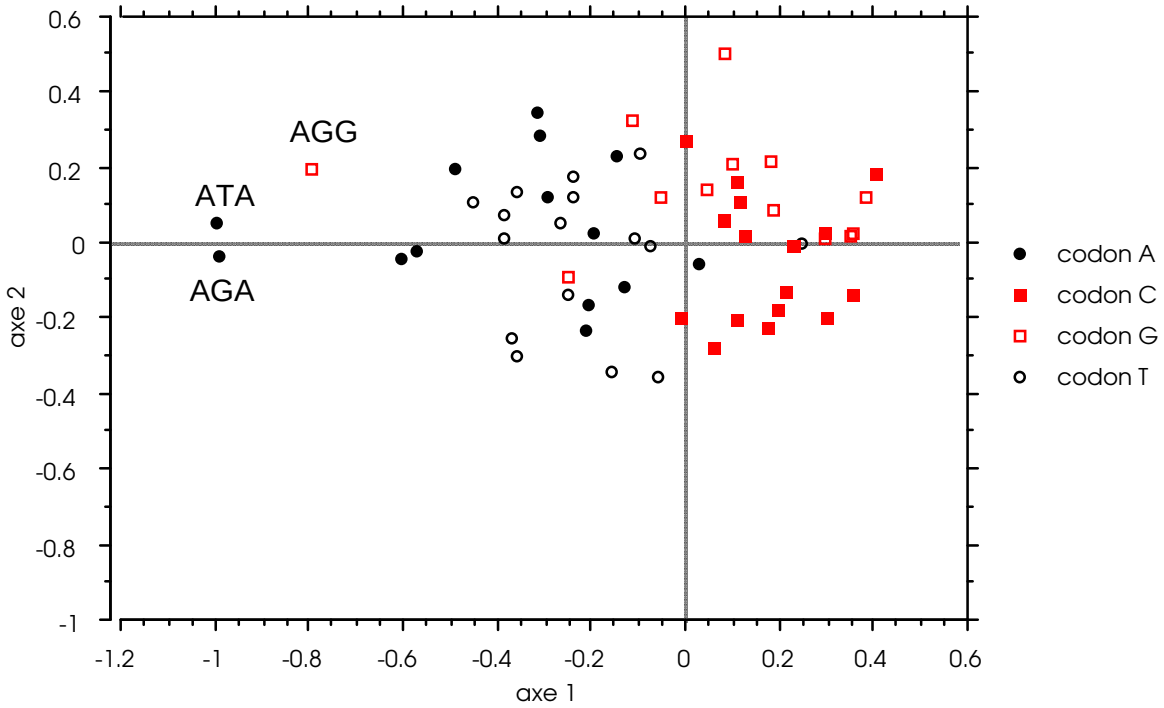
Fig. 3.23 : Pages suivantes, AFC intra-espèces pour les quatre espèces considérées.
En haut, les gènes. Les différentes classes sont représentées : non-HT : gènes natifs ; HT : gènes acquis récemment (Pour *E. coli*, set1 : transferts après la séparation des O157:H7 et set3 : transferts avant la séparation des O157:H7) ; IS : Séquences d'insertion ; Ph : phages.
En bas, les codons. Les codons sont éclatés en fonction de la nature de la base en 3ème position (A,T,C,G). Les deux graphes sont superposables ce qui permet d'identifier les codons responsables des regroupements de gènes. Le pourcentage de variance expliquée par les axes est indiqué entre parenthèses.

Escherichia coli

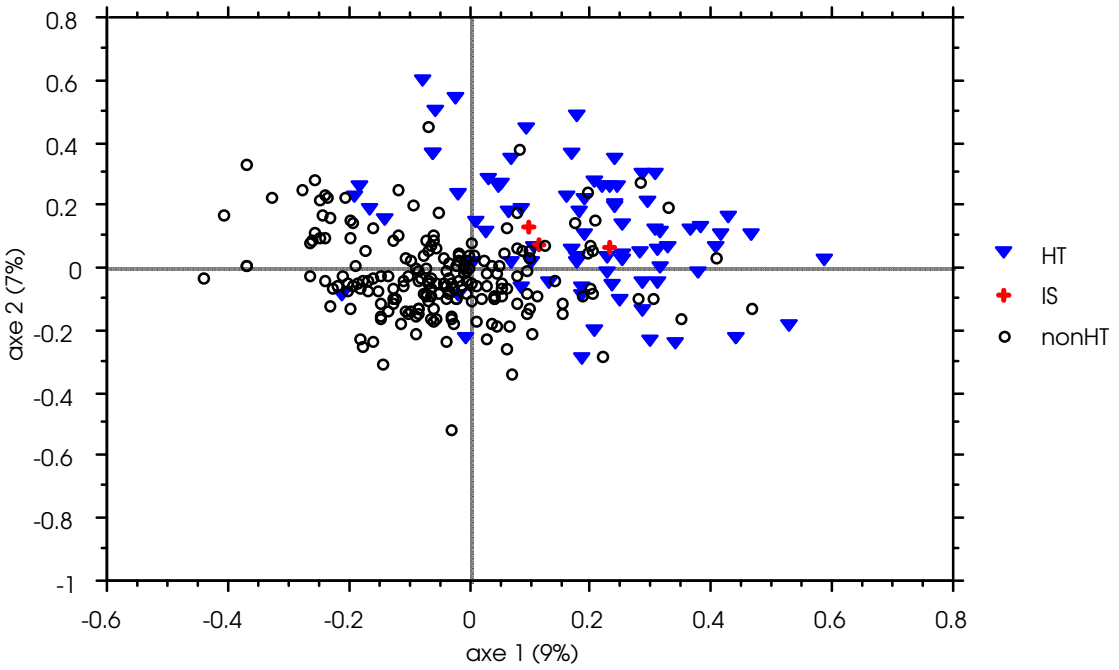


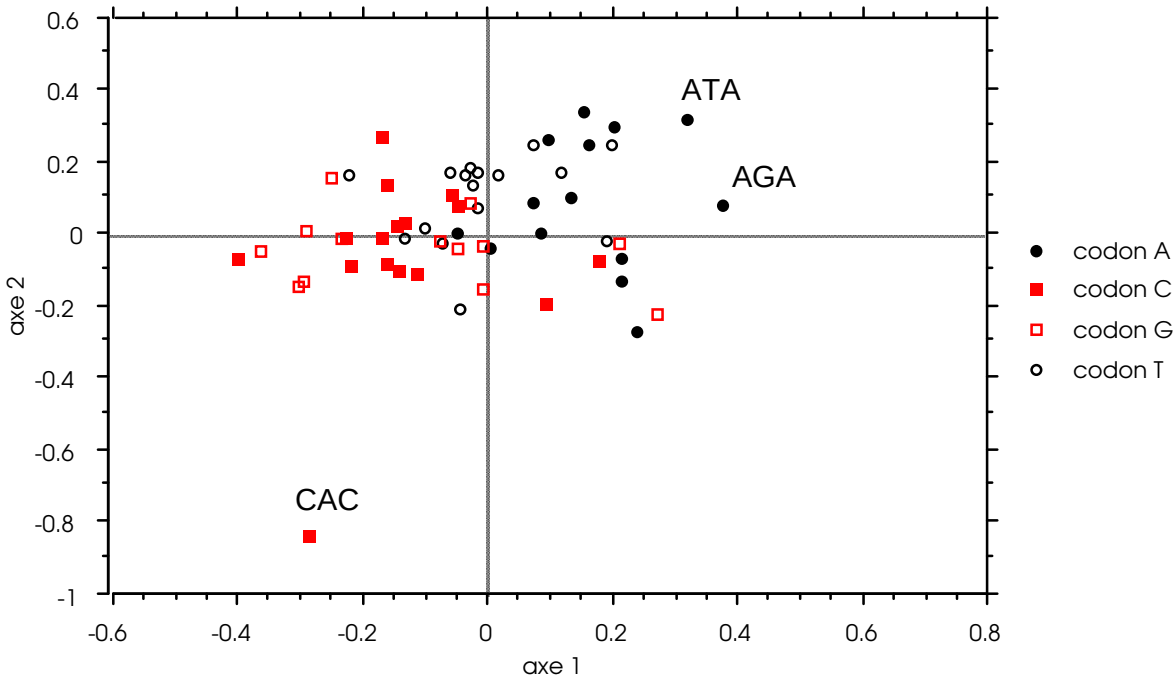
Salmonella enterica



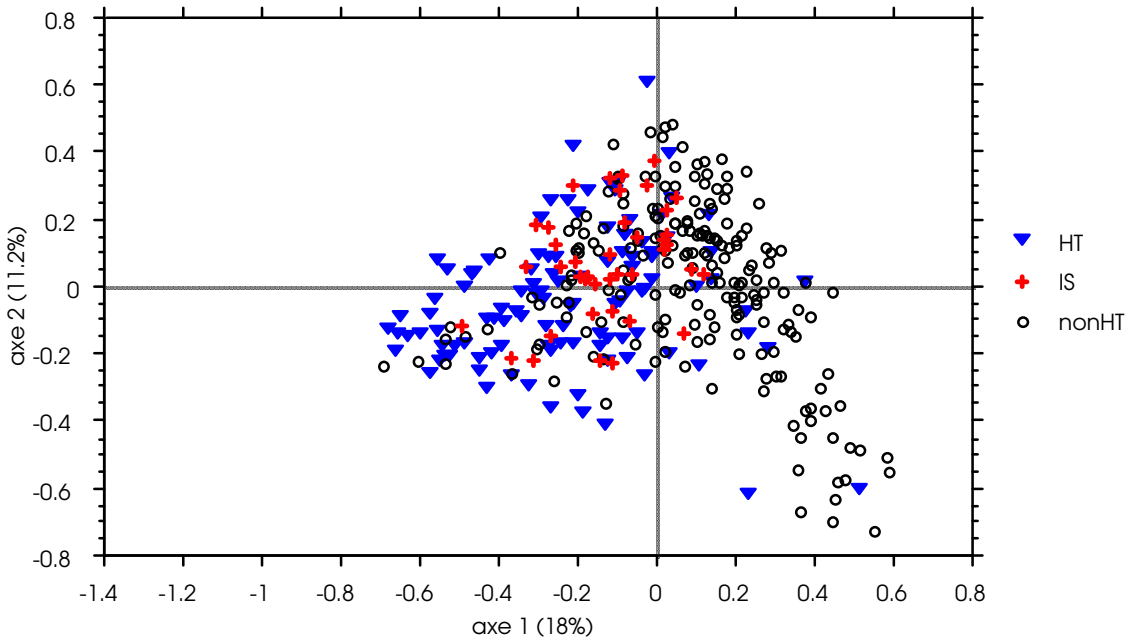


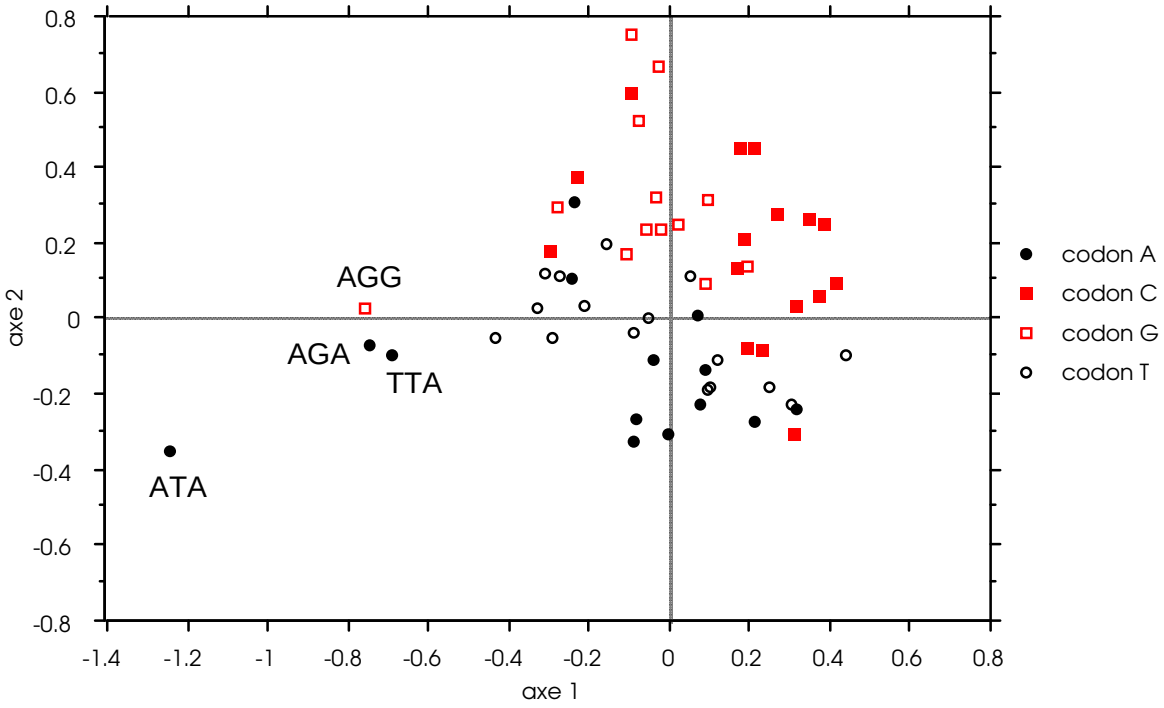
Helicobacter pylori





Streptococcus pneumoniae





Chez *E. coli* et *Salmonella*, ce premier axe correspond essentiellement à une séparation des codons se terminant par A et T des codons se terminant par G et C. Chez *Helicobacter*, on observe la même tendance même si ce sont les codons se terminant par A qui semblent être plus fortement opposés aux codons se terminant par C. Chez *Streptococcus*, la séparation A+T3/G+C3 se fait plutôt sur le deuxième axe. Dans toutes ces analyses, on peut remarquer que le codon ATA (codant pour l'isoleucine) est systématiquement placé à un des extrêmes du premier axe, la plupart du temps en compagnie du codon AGA (Arginine). Le codon AGG (Arginine) est également proche de ces codons dans les AFC correspondant à *E. coli*, *Salmonella* et *Streptococcus* (en compagnie du codon TTA-Leucine, dans ce cas). Ceci suggère que ces codons sont particulièrement responsables de la séparation des gènes transférés et natifs. Si l'on observe les fréquences de ces codons dans les différents groupes de gènes, l'on s'aperçoit que pour l'isoleucine chez toutes les espèces et l'arginine chez *Escherichia* et *Salmonella*, les gènes natifs montrent un évitement de ces codons alors que les gènes transférés montrent des fréquences plus proches de l'équiprobabilité. Pour l'arginine, les gènes acquis récemment par *Helicobacter* et *Streptococcus* montrent un net biais vers le codon AGA, qui est absent des gènes natifs.

		Helicobacter			*Salmonella*				*Escherichia*					*Streptococcus*		
										nouveaux	anciens					
		Natifs	HT	IS	Natifs	HT	IS	Phages	Natifs	HT	HT	IS	Phages	Natifs	HT	IS
I	**ATA**	**0,12**	**0,26**	**0,27**	**0,08**	**0,23**	**0,29**	**0,12**	**0,06**	**0,32**	**0,25**	**0,23**	**0,14**	**0,08**	**0,25**	**0,12**
	ATT	0,50	0,50	0,36	0,49	0,48	0,33	0,47	0,51	0,37	0,46	0,34	0,46	0,54	0,57	0,48
	ATC	0,38	0,24	0,37	0,43	0,29	0,38	0,41	0,43	0,31	0,29	0,43	0,40	0,38	0,18	0,39
R	**AGA**	**0,26**	**0,45**	**0,58**	**0,03**	**0,14**	**0,16**	**0,07**	**0,03**	**0,18**	**0,15**	**0,08**	**0,09**	**0,14**	**0,36**	**0,25**
	AGG	**0,25**	**0,18**	**0,21**	**0,02**	**0,10**	**0,16**	**0,05**	**0,02**	**0,16**	**0,08**	**0,07**	**0,07**	**0,04**	**0,11**	**0,07**
	CGA	0,07	0,07	0,04	0,06	0,11	0,19	0,08	0,06	0,10	0,11	0,12	0,08	0,11	0,11	0,20
	CGT	0,14	0,14	0,06	0,35	0,26	0,24	0,30	0,39	0,17	0,26	0,30	0,30	0,50	0,28	0,27
	CGC	0,25	0,14	0,06	0,43	0,23	0,13	0,36	0,41	0,23	0,25	0,26	0,26	0,17	0,10	0,16
	CGG	0,03	0,02	0,04	0,11	0,15	0,11	0,13	0,09	0,17	0,15	0,16	0,20	0,04	0,04	0,07

Tableau 3.2 : Fréquences relative des codons synonymes de l'Isoleucine (**I**) et de l'Arginine (**R**) pour les différentes classes de gènes dans les quatres espèces. Les codons en gras sont ceux identifiés comme étant particulièrement discriminants dans l'AFC.

Comme nous l'avons déjà mentionné plus haut, la presque totalité des gènes détectés comme transférés horizontalement par la méthode utilisée n'ont pas de fonction connue. Par contre, certains gènes non détectés par notre méthode se trouvent dans le nuage des gènes

transférés horizontalement. Les fonctions de ces gènes laissent supposer qu'il s'agit effectivement de gènes transférés. Ainsi, on trouve beaucoup de protéines membranaires liées à la virulence, à des systèmes de sécrétion. Chez *Streptococcus* et *Helicobacter*, on trouve également des enzymes de restriction et de régulateurs de transcription. De manière surprenante, nous avons identifié parmi ces gènes chez *H. pylori* une protéine ribosomale (RPS14). Celle ci a été décrite sur des arguments phylogénétiques comme ayant subi de nombreux transferts, notamment chez les proteobactéries et pourrait être en relation avec un mécanisme de résistance à certains antibiotiques (Brochier, *et al.*, 2000).

3.3.2.4 AFC sur les gènes de quatre espèces

La fig. 3.24 représente les deux premiers axes de l'AFC sur les gènes des quatre espèces. La part de variance expliquée par les axes 1 et 2 est respectivement de 22,98 % et 7,29 %. Les ellipses représentent 90 % des points d'une catégorie, et un test de MANOVA montre que chacune des catégories présentées est significativement distincte des autres ($p < 10^{-4}$). La figure est décomposée en quatre parties superposables pour une plus grande lisibilité. Le premier axe est surtout déterminé par la richesse en A+T des codons (voir fig. 3.24D). La fig. 3.24A représente les ellipses correspondant aux gènes natifs. Le centre de gravité de chaque ellipse est indiqué par un point. Les deux autres parties (fig 3.24B et 3.24C) représentent respectivement les gènes transférés horizontalement et les IS. Les phages ne sont pas représentés du fait de l'absence de données pour *Streptococcus* et *Helicobacter*. Les flèches représentent le déplacement du centre de gravité des gènes transférés et des IS par rapport à celui des gènes natifs. On remarque que pour les gènes transférés horizontalement, le déplacement du centre de gravité se fait essentiellement sur le premier axe, en direction des codons riches en A+T, et que ces gènes montrent une moins grande disparité que les gènes natifs au niveau de l'axe 2. Le décalage vers l'A+T des gènes transférés est d'autant plus important que les gènes natifs sont riches en G+C. Le déplacement du centre de gravité des IS se fait de manière assez analogue sauf pour *E. coli*, où le déplacement se fait principalement sur l'axe 2. Il est intéressant de noter que pour *E. coli* et *S. enterica*, les gènes natifs d'une part et les gènes transférés horizontalement d'autre part occupent des positions proches dans l'AFC.

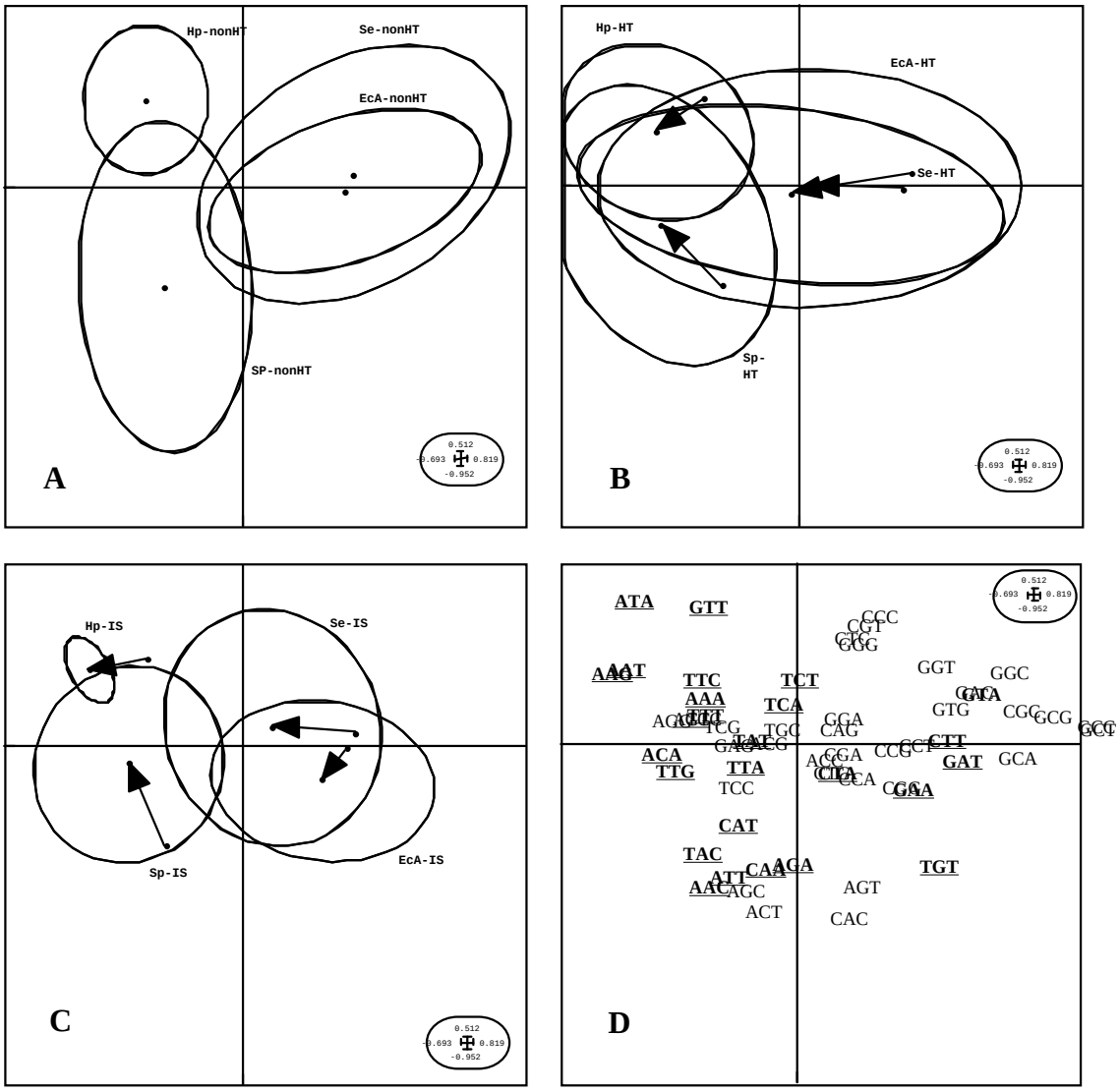


Fig. 3.24 : AFC inter-espèce. Les quatre parties de la figure sont superposables. Les ellipses représentent 90% des points d'une classe. Hp :*Helicobacter pylori* ; Se : *Salmonella enterica* ; EcA : *Escherichia coli* ; Sp : *Streptococcus pneumoniae*. Non_HT : gènes natifs ; HT : gènes acquis récemment ; IS : séquences d'insertion. En bas à droite, les codons correspondant. Les codons relativement riches en A+T (contenant au moins deux A ou T) sont soulignés. Les flèches représentent le déplacement du centre de l'ellipse de gènes transférés et des IS par rapport au centre de l'ellipse des gènes natifs.

3.3.2.5 La composition en bases des gènes transférés horizontalement, phages et IS

Les gènes récemment acquis et perdus montrent, pour tous les génomes étudiés des taux de G+C en troisième position des codons très significativement inférieurs au reste du génome ($p < 0{,}0001$, test de Mann-Whitney) (voir plus loin, fig 3.25). Ceci est également vrai, quoique moins marqué pour les première et deuxième positions des codons. Ce résultat est particulièrement surprenant pour des espèces comme *Streptococcus* et *Helicobacter* dont les génomes ont des taux de G+C3 relativement bas (respectivement 35 % et 41 %). Ceci signifie que quelle que soit la richesse en A+T du génome, les gènes transférés ont tendance à être plus riches en A+T que leur génome hôte. De manière intéressante, dans les génomes où les effectifs de gènes ayant été perdus sont suffisants (les deux *Salmonella* et les *E. coli* K12), il existe également un forte tendance ($p < 0{,}0001$, test de Mann-Whitney) de ces gènes à être plus riches en A+T que le reste du génome. Ceci suggère un « turn-over » plus important pour les gènes riches en A+T dans ces génomes. Chez *E. coli* O157:H7, Il est intéressant de noter que les gènes transférés les plus anciens montrent des taux de G+C légèrement supérieurs aux gènes présents dans le génome depuis moins longtemps, ce qui peut être le témoignage du processus d'amélioration décrit par Lawrence et Ochman (Lawrence et Ochman, 1997).

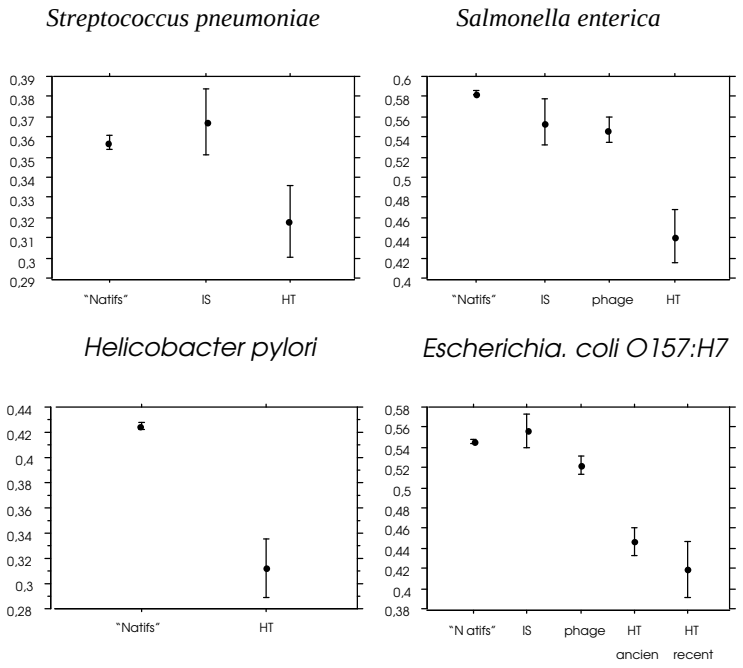


Fig.3.25 : Taux de G+C3 moyen des différentes classes de gènes identifiées dans cette analyse. Les barres représentent 95 % d'intervalle de confiance. HT : gènes transférés horizontalement (selon notre méthode) ; HT récent : gènes acquis depuis la séparation des deux *E. coli* 0157:H7; HT anciens : gènes acquis avant la séparation des deux O157:H7 ; IS : gènes annotés comme appartenant à un élément transposable bactérien ; phages : gènes de bactériophages ; natifs : gènes n'appartenant à aucune de ces classes.

Rocha et Danchin (2002) ont récemment mis en évidence que les éléments parasites des génomes étaient couramment biaisés vers les

nucléotides A+T. Nos résultats montrent que c'est également le cas des gènes transférés, et ce quelle que soit la richesse en A+T du génome hôte. Les mécanismes de transferts des gènes impliquent souvent des systèmes tels que les IS et les phages, et l'on peut supposer que le fait d'utiliser ces moyens de transport biaise la composition en bases de ces gènes. Cependant, les résultats de l'AFC suggèrent que les IS et plus particulièrement les phages sont beaucoup moins biaisés au niveau de leur composition en A+T. La fig. 3.25 montre les taux de G+C3 des gènes des différentes classes étudiées ici dans les quatre espèces. Les mêmes tendances sont observées pour le taux de G+C aux autres positions (résultats non présentés). Seuls les gènes natifs et transférés sont présentés pour *Helicobacter* car les génomes des deux souches ne contiennent pas de phages annotés comme tels et seulement un nombre très faible d'IS. De même, *Streptococcus* ne montre pas de séquences de phages annotées. Ces résultats contredisent l'hypothèse précédemment formulée que les gènes transférés horizontalement pourraient adopter la composition en base de leurs « moyens de transport » car ces derniers semblent présenter une richesse en A+T moindre. Il faut noter cependant que les phages considérés ici sont des phages présents dans la séquence complète des génomes et donc qu' il s' agit en cela de phage tempérés. Rocha et Danchin (2002) ont en effet montré que les phages virulents avaient une tendance a être plus riches en A+T que les tempérés.

3.3.2.6 Sélection agissant sur les différentes classes de gènes

Sueoka (Sueoka, 1988) a proposé un moyen de mesurer l'importance de la sélection s'exerçant sur un gène qu'il a appelé « Relative Neutrality Plot ». Cette méthode consiste à tracer un graphique représentant le taux de G+C aux positions contraintes au niveau de la protéine (positions 1 et 2 des codons) en fonction de la position la moins contrainte par la sélection, la position 3. Bien que la troisième position des codons ne soit pas purement neutre, elle est celle qui présente la plus grande variabilité en contenu en bases au sein des génomes. Elle présente également la meilleure corrélation entre son contenu en G+C et celui des régions non codantes voisines. Enfin, l'effet de la sélection traductionnelle à cette position est faible en comparaison de la sélection au niveau de la protéine. De ce fait, elle est la position qui reflète le mieux les contraintes mutationnelles qui s'exercent sur le gène (Sueoka, 1995; Sueoka, 1999). La pente attendue de la corrélation linéaire calculée entre les paramètres est égale à 1 si les séquences ne subissent aucune contrainte sélective, et est d'autant plus faible que la sélection est forte. Sueoka (Sueoka, 1999) a appliqué cette méthode pour quantifier

l'influence relative de la mutation et de la sélection sur l'évolution des génomes bactériens. Dans notre cas, il est intéressant d'étudier ces corrélations en fonction des différentes classes de gènes, afin de savoir si certaines sont plus soumises à sélection que d'autres. Les résultats présentés fig. 3.26 représentent le taux de G+C en première et deuxième positions en fonction du taux de G+C3 pour *Escherichia coli* O157:H7. Les mêmes tendances ont été trouvées chez *Salmonella*, *Streptococcus* et *Helicobacter*. Comme attendu, les gènes natifs montrent une corrélation avec une pente assez faible entre ces deux paramètres (0,241 ; R^2=0,212). De manière intéressante, ce sont les gènes acquis le plus récemment qui présentent la pente la plus forte (0,568 ; R^2=0,446), suivis des transferts plus anciens (0,451 ; R^2=0,553), ce qui témoigne peut-être de l'amélioration de ces derniers (au sens de Lawrence et Ochman, 1997). Ces fortes pentes témoignent du fait que le contenu en base de ces gènes est essentiellement déterminé par des pressions de mutation. Les phages montrent une pente (0,3 ; R^2=0,392) beaucoup plus proche des gènes natifs ce qui témoigne d'une pression de sélection plus importante que sur les gènes transférés récemment. Pour comparaison, nous avons également calculé la pente de la corrélation pour des gènes de plasmides d'*E. coli* extraits de Genbank Release 130 (Benson, *et al.*, 2002) : le coefficient (0,288 ; R^2=0,301) est proche de celui observé pour les phages.

Les gènes des éléments transposables (IS) montrent une absence totale de corrélation (R^2=0,001) ce qui est particulièrement surprenant. Cette tendance est retrouvée également chez *Salmonella* et *Streptococcus*, ce qui suggère que ce résultat n'est ni un artefact, ni un cas particulier à *E. coli* O157:H7. Il semble ainsi que la composition en bases des première et deuxième bases des codons soit indépendante de la composition de la troisième chez les éléments transposables de ces trois espèces.

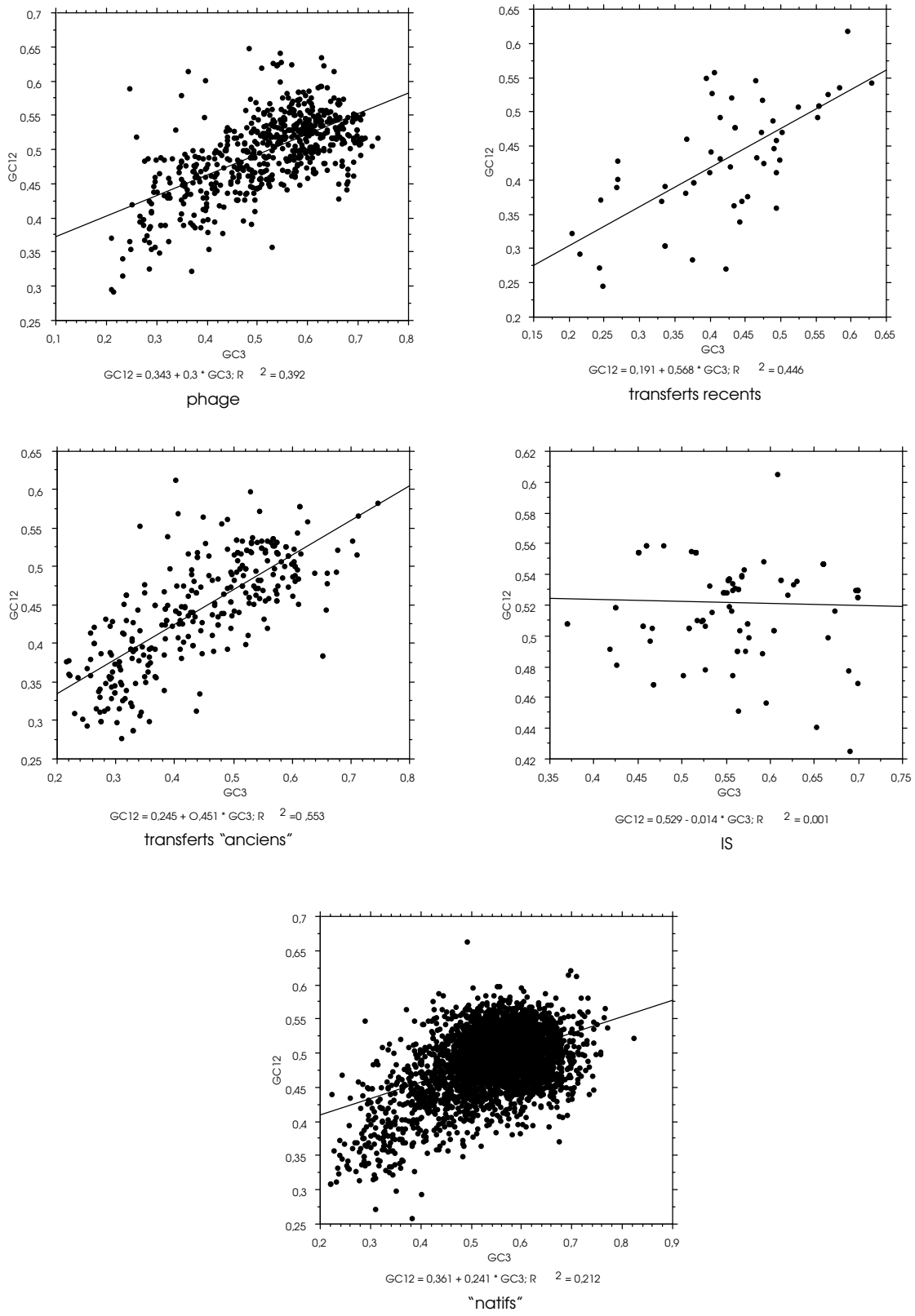


Fig.3.26 : « Relative Neutrality Plot » pour les différentes classes de gènes chez *E. coli* O157 :H7. On observe exactement les mêmes tendances dans les autres espèces. Le coefficient de corrélation est particulièrement élevé pour les gènes acquis récemment.

3.3.3 ***Discussion***

3.3.3.1 Le terminus, un site préferentiel d'insertion ?

D'après le test de χ^2 effectué, plusieurs génomes montrent une sur-représentation significative des gènes acquis récemment dans une grande région située autour du terminus. Nous avons vu dans le chapitre précédent que la présence de sites *ter* pouvait bloquer la fourche de réplication dans cette région et ainsi potentiellement augmenter la probabilité de recombinaison avec un ADN exogène du fait de la présence d'ADN simple brin. Cependant, cette tendance n'est pas observée dans tous les génomes. De plus, la distribution de ces gènes dans le génome d'*E. coli* O157:H7 présentée fig. 3.22 suggère que c'est une région relativement proche du terminus plutôt que le terminus lui même qui est sujette à de nombreuses insertions. Or, les évenements d'insertion des gènes dans le génome ne sont pas indépendants les uns des autres. C'est un phénomène de ce type qui est à l'origine de la formation d'îlots de pathogénicité (Hacker et Kaper, 2000). Ainsi, l'insertion d'un îlot à un site dans le génome va augmenter la probabilité d'insertion d'un autre îlot au même site du fait de la recombinaison homologue entre les répétitions directes qui les encadrent et donc provoquer l'accumulation de gènes transférés à ce site. La nature contingente de ce processus d'insertion peut provoquer une corrélation artificielle avec la position du terminus. Perna *et al.* (Perna, *et al.*, 2001) ont montré que chez *E. coli* O157:H7, la taille des îlots a tendance à être beaucoup plus importante que chez *E. coli* K12. La fig. 3.22 suggère que la région proximale au terminus pourrait contenir un regroupement d'îlots. Le regroupement de ces gènes acquis récemment pourrait donc ne pas être dû à la proximité du terminus en elle même.

3.3.3.2 La richesse en A+T des gènes transférés horizontalement

La tendance des gènes transférés horizontalement à être riches en A+T a déjà été notée par de nombreux auteurs (Medigue, *et al.*, 1991; Syvanen, 1994; Lawrence et Ochman, 1997). Nous avons cependant montré que cette tendance est vraie même pour des génomes

relativement (*H. pylori*) ou très (*S. pneumoniae*) riches en A+T. Ces caractéristiques communes tendent à remettre en question certaines idées sur les transferts horizontaux. Notamment, Lawrence et Ochman (Lawrence et Ochman, 1997) supposent, lorsqu'ils détectent les gènes récemment acquis et qu'ils calculent leur degré d'amélioration, que ces gènes sont adaptés au contexte génomique d'une espèce éloignée. Nos résultats suggèrent, selon cette hypothèse, que les génomes donneurs ont une forte tendance à être plus riches en A+T que les génomes accepteurs, ou bien qu'il éxiste un biais au niveau de l'entrée et/ou de l'ADN exogène dans le génome. Ceci pourrait s'expliquer par une barrière physique liée à la pénétration dans la cellule, comme par exemple une tendance des enzymes de restriction à posséder des sites de reconnaissance relativement riches en G+C. Nous avons en effet calculé que les sites listés dans la base d'enzyme de restriction REBASE (Roberts et Macelis, 2000) présentent un taux moyen de G+C supérieur à 70 %, après élimination de la redondance, mais ceci pourrait être lié à un biais introduit par la sélection d'enzymes d'intérêt industriel. Cependant, on comprend mal pourquoi le biais des gènes transférés se ferait toujours relativement au génome hôte. En effet, si les gènes transférés ont une forte tendance à être plus riches en A+T que leur génome hôte, ceux présents chez les bactéries riches en A+T sont plus riches en A+T que ceux présents chez les bactéries ayant un taux de G+C moyen. L'hypothèse d'un crible dû aux enzymes de restriction paraît donc peu probable.

3.3.3.3 Les gènes récemment acquis portent-t-ils la marque d'hôtes antérieurs ?

Une autre caractéristique de ces gènes est leur tendance à adopter une composition en bases, et probablement en acides aminés, principalement déterminée par la mutation, comme le suggèrent les « Relative Neutrality Plot ». Ce biais ne semble pas avoir d'équivalent dans les classes de gènes que nous avons étudiées et les pentes observées entre le G+C12 et le G+C3 s'apparentent plus à celles rapportées par Sueoka (Sueoka, 1999) pour la corrélation entre le G+C intergénique et le G+C3 (de l'ordre de 0,7). On pourrait imaginer que ces séquences représentent des erreurs de prédiction de gènes et ne sont pas exprimées. Cependant, bien que la plupart des gènes détectés comme ayant été acquis récemment n'ont ni homologue, ni fonction connue, leur usage du code s'apparente fortement à ceux de gènes de la virulence, de résistance, de protéines de sécrétion etc... (voir section 3.3.2.3) De plus, Alimi *et al.* (Alimi, *et al.*, 2000) ont montré qu'il est probable que les gènes orphelins prédits chez *E. coli* soient effectivement transcrits.

Ainsi, il semble que quelle que soit l'espèce dans laquelle on les trouve, les gènes transmis horizontalement portent les traces d'un fort biais mutationnel vers A+T, dont l'intensité est bien supérieure à celle observée non seulement chez les gènes natifs mais également chez d'autres séquences parasites des génomes bactériens comme les plasmides, les IS ou les phages. Ceci suggère que ces gènes forment bien, comme l'avaient implicitement supposé Médigue *et al.* (Medigue, *et al.*, 1991), une classe de gènes cohérente non seulement pour leur taux de G+C, mais également pour les contraintes fonctionnelles et mutationnelles qu'ils subissent.

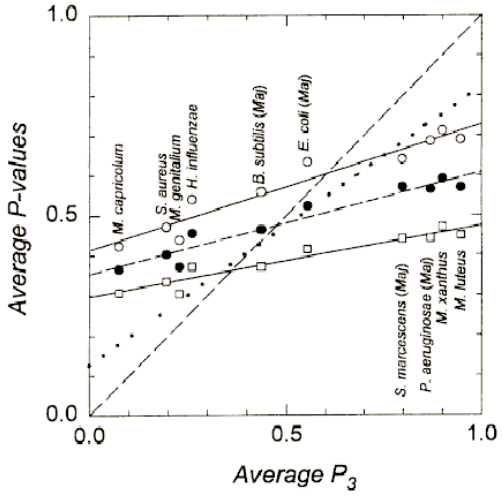


Fig. 3.27 : « Relative Neutrality Plot » sur un échantillonnage représentatif des espèces bactériennes en terme de taux de G+C du génome. Le graphe représente le taux de G+C moyen en première position (G+C1, carrés blancs), en deuxième position (G+C2, points blancs) et la moyenne pour ces deux positions (G+C12, points noirs) en fonction du taux de G+C3. Extrait de Sueoka, 1999.

Sueoka (Sueoka, 1999) a calculé la pente de la corrélation des différentes positions des codons au G+C3 dans un échantillon représentatif d'espèces bactériennes. Il montre que la relation entre G+C3 moyen et G+C12 moyen est constante dans tous les génomes bactériens, et que le coefficient de corrélation qui lie ces deux facteurs est de l'ordre de 0,25. Il en résulte que la corrélation attendue entre ces facteurs pour des gènes acquis d'espèces bactériennes prises au hasard est de 0,25. Nous avons confirmé cette prédiction par des tirages aléatoires de gènes bactériens dans la base de données GenBank Release 130 (Benson, *et al.*, 2002). Les pentes calculées sont proches de 0,3 (données non présentées).

Lawrence et Ochman (Lawrence et Ochman, 1997) ont modélisé la dynamique d'amélioration d'un gène apparaissant dans un génome. Selon leur modèle, la position non contrainte (la troisième position des codons) est celle dont l'adaptation au biais mutationnel du nouvel hôte se fait le plus rapidement, la composition des autres positions et particulièrement de la deuxième position des codons variant très peu. On pourrait imaginer que la forte pente pour les gènes transférés soit due au processus d'amélioration qui doit modifier le taux de G+C des gènes de manières différentes pour les trois positions des codons.

Cependant, la divergence des gènes natifs entre les deux souches d'*E. coli* O157:H7 est si faible (les séquences sont quasiment toutes identiques au niveau nucléique) qu'il est difficile d'envisager que des gènes acquis après leur séparation aient eu le temps de subir un tel biais. On ne peut donc pas attribuer le fort coefficient observé chez les gènes acquis récemment à la pression de mutation qu'ils subissent dans leur nouvel environnement génomique. Par contre, les gènes acquis depuis plus longtemps (« HT anciens », fig 3.26) montrent une pente plus faible que ceux acquis plus récemment, ce qui peut témoigner d'un processus d'amélioration en marche.

Il semble donc que, de même que les IS, les phages ou les plasmides ne présentent pas une composition en base héritée d'un hôte quelconque (Rocha et Danchin, 2002), les gènes acquis récemment n'aient pas les caractéristiques attendues de gènes adaptés à un hôte antérieur. Cependant, ces gènes se démarquent également des IS, phages et plasmides ce qui suggère que leur composition en bases n'est pas un conséquence du fait qu'ils peuvent utiliser ces parasites du génome pour se déplacer. Les différences importantes de ces gènes au niveau de leur composition en bases suggèrent que les méthodes de prédiction des gènes transférés horizontalement identifient effectivement des gènes dont la présence dans le génome est récente. Cependant, il se pourrait que les raisons de ces différences et donc de leur identification comme gènes transférés ne soient pas celles couramment invoquées. Si nos résultats soutiennent l'existence d'un processus d'amélioration comme celui décrit par Lawrence et Ochman (Lawrence et Ochman, 1997), il semble cependant que pour la plupart des gènes, cette adaptation au nouvel environnement génomique ne se fasse pas à partir d'un usage du code typique d'une autre espèce bactérienne.

Les pressions responsables de la richesse en A+T des gènes transférés restent obscures. Rocha et Danchin (Rocha et Danchin, 2002) ont suggéré que certaines séquences parasites des génomes pouvaient avoir un intérêt à s'enrichir en A+T, du fait de la plus grande disponibilité des ATP dans la cellule. Cependant, si ce modèle pourrait expliquer la richesse en A+T de gènes transférés égoïstes, il faut y ajouter une atténuation de cette pression avec le temps, puisque plus les gènes persistent dans le génome, plus ils s'appauvrissent en A+T. Reste l'hypothèse d'un biais d'incorporation des gènes riches en A+T dans les génomes au niveau de la pénétration dans la cellule ou encore de l'incorporation au génome. Comme nous l'avons déjà noté, il semble que les enzymes de restriction dont le site de coupure est connu présentent une préférence statistique pour les sites riches en G+C. Cependant, la recherche

des sites de restriction ne fait probablement pas l'objet d'un criblage aléatoire et des intérêts industriels peuvent biaiser ce résultat. En outre, cette hypothèse est difficilement envisageable dans le cas des génomes riches en A+T. Il est particulièrement remarquable qu'une très forte proportion des gènes identifiés comme transférés horizontalement n'aient aucun homologue connu chez aucune autre espèce. Comme le notent Alimi *et al.* (Alimi*, et al.*, 2000), malgré l'accumulation de séquences dans un spectre d'espèces de plus en plus vaste, chaque nouveau génome séquencé présente un forte proportion de gènes non caracterisés, dont un nombre significatif sont des gènes orphelins stricts. Il semble que les gènes transférés appartiennent préférentiellement à cette dernière classe, ce qui rend la compréhension de leur évolution particulièrement délicate.

Discussion générale et conclusion

4 Discussion générale et conclusion

La reconstruction de l'histoire et des mécanismes évolutifs à l'œuvre chez les procaryotes constitue un défi pour les biologistes. Certes, nous commençons à disposer de très grandes quantités de séquences, mais notre vision de la diversité de ces organismes est encore extrêmement lacunaire. Des données actuellement disponibles certains auteurs ont déduit l'image d'un monde procaryote complètement chimérique, où l'abondance des transferts horizontaux serait telle que les génomes ne seraient que des vecteurs transitoires de gènes, et que tenter de reconstruire la phylogénie des espèces serait vain. Cependant l'étude attentive des données et des résultats qui ont conduit à ce constat montre un amalgame de plusieurs catégories de faits dont la connexion n'est pas forcément très claire.

D'un coté, le contenu en gènes des différentes souches d'une même espèce peut être extrèmement variable. Par exemple, si le génome d'*E. coli* K12 contient de l'ordre de 4600 gènes, ceux des souches pathogènes O157:H7 peuvent en avoir plus de 5600. L'analyse du contenu en G+C des gènes des génomes de ces bactéries révèle qu'il existe une variabilité importante entre les gènes, et que, parmi les plus atypiques d'entre eux en termes de composition, se trouvent des gènes liés à des contraintes sélectives fortes (comme la résistance aux antibiotiques, la virulence etc...). Les différences de contenu de ces gènes sont interprétées comme l'indice qu'ils proviennent d'espèces éloignées. Cependant, une majorité de ces gènes ont des fonctions inconnues et correspondent même à des « orphelins », qu'on ne retrouve dans aucune autre espèce connue (voir section 1.7).

D'un autre coté, un certain nombre d'études phylogénétiques ont montré que des transferts pouvaient avoir lieu pour des gènes impliqués dans des fonctions cellulaires très diverses, jusque dans les plus fondamentales d'entre elles (voir par exemple Brochier*, et al.*, 2000), et ce même entre organismes éloignés. Un petit nombre de ces cas spectaculaires ont été bien décrits et montrent indubitablement que les transferts horizontaux sont un fait réel dans le monde procaryote et permettent l'exploration de nouveaux milieux. En effet, l'étude attentive de ces cas révèle souvent un lien avec un avantage sélectif fort : on peut citer par exemple l'acquisition par plusieurs bactéries d'une Isoleucyl-ARNt synthétase eucaryote

ayant probablement conféré à ces premières des propriétés de résistance à certains antibiotiques, ou encore le transfert de gènes d'ATP/ADP translocase eucaryotes à des parasites intracellulaires bactériens, ce qui leur permet d'utiliser l'ATP de leur hôte (voir Koonin, *et al.*, 2001 également pour d'autres exemples). Des transferts entre bactéries d'un gène codant pour une protéine ribosomique ont également été identifiés, probablement en relation avec la résistance à un antibiotique (Brochier, *et al.*, 2000). D'autre part, de très nombreuses phylogénies présentent des incongruences marquées. Dans ces cas, aucun transfert horizontal en particulier n'est identifiable. Cependant, ces topologies aberrantes sont, par défaut, interprétées en terme de transferts horizontaux (Jain, *et al.*, 1999; Nesbo, *et al.*, 2001; Zhaxybayeva et Gogarten, 2002). L'abondance des transferts observés au niveau des souches bactériennes a fait du transfert horizontal l'hypothèse la plus parcimonieuse pour expliquer les incongruences phylogénétiques. Or, les méthodes phylogénétiques, surtout à l'echelle d'un groupe aussi vaste et diversifié que les bactéries, ne sont pas exemptes d'artefacts. Peut-on à ce point faire confiance aux méthodes phylogénétiques et les deux catégories de méthodes (phylogénétique et composition des gènes) observent-elles vraiment le même phénomène ?

Nous avons montré dans ce travail qu'il est possible d'extraire des familles de gènes, des informations congruentes sur la phylogénie des bactéries. Ceci requiert de mettre au point des méthodes de recherche de la congruence des données. D'autres travaux, par des méthodologies très différentes obtiennent des résultats très semblables, au moins en ce qui concerne la phylogénie des bactéries (Brochier, *et al.*, 2002). Ces résultats relativisent l'idée selon laquelle la métaphore de l'arbre serait inapropriée pour représenter l'histoire des procaryotes (voir section 1.8). De plus, il est assez remarquable que lorsque l'on compare les topologies obtenues pour différentes familles de gènes, ce sont les arbres qui contiennent le plus d'espèces qui tendent à être les plus congruents entre eux (voir section 2.3.2.2.). Ceci suggère qu'il est possible que la majorité des arbres incongruents le soient parce que les méthodes de reconstruction sont incapables de reconstruire leur histoire. Plusieurs travaux ont en effet montré qu'un faible échantillonnage taxonomique pouvait avoir des conséquences désastreuses sur la reconstruction phylogénétique (Lecointre, *et al.*, 1993; Philippe et Douzery, 1994; Adachi et Hasegawa, 1996). De ce point de vue, il est symptomatique que les études basées sur la méthode de « Likelihood Mapping » utilisant des quartets (arbres à quatre espèces) (voir section 2.2.2) soient celles qui observent le plus d'incongruences entre les données et concluent à des transferts extensifs (Nesbo, *et al.*, 2001; Zhaxybayeva et Gogarten,

2002). Jain *et al.* (Jain, *et al.*, 1999), en utilisant des comparaisons de topologies, ont remarqué que les arbres reconstruits à partir de gènes impliqués dans des fonctions cellulaires essentielles comme la réplication, la transcription et la traduction tendent à présenter moins d'incongruences (voir aussi section 2.2.1). Ils suggèrent, probablement à juste titre, qu'un transfert est d'autant moins susceptible de fonctionner que la protéine codée par le gène a des interactions multiples et complexes avec les autres protéines de la cellules. Cependant, si un telle protéine est soumise à des pressions de sélection qui rendent improbable son transfert, sa séquence doit également être plus contrainte qu'une autre, et son taux d'évolution moindre. Ainsi, le résultat de Jain *et al.* (Jain, *et al.*, 1999) peut également s'interpréter comme suit : nous savons mieux reconstruire la phylogénie des protéines ayant des interactions multiples, du fait des contraintes particulières qui s'exercent sur elles.

Nous présentons également des résultats qui suggèrent que les gènes qui sont responsables des différences remarquables de contenu en gènes des génomes entre souches sont atypiques à bien des égards. La forte tendance à une richesse en A+T par rapport au génome hôte tend à montrer que ces gènes ne présentent pas des caractéristiques attribuables à un quelconque hôte antérieur, et leur composition aussi bien en bases qu'en acides aminés semble avant tout être déterminée par des pressions (probablement mutationelles) s'exerçant au niveau de la séquence nucléique. Cependant, nous avons également montré que la seule richesse en A+T ne pouvait pas constituer un critère de détection des gènes acquis horizontalement du fait d'une structuration intrinsèque des génomes, probablement liée à la réplication. L'écrasante majorité des gènes qui ont effectivement été acquis récemment n'appartiennent pas à des familles pour lesquelles il est envisageable de reconstruire une phylogénie : la plupart sont rarement ou pas du tout représentées dans d'autres génomes. Si cette observation met l'accent sur les lacunes de notre perception du monde procaryote, elle suggère surtout que ces gènes appartiennent à une catégorie à part, qu'il est hasardeux de rapprocher de celle des gènes dont la fonction est caractérisée et qui sont utilisés pour reconstruire des phylogénies.

Ainsi, les procaryotes semblent échanger de l'ADN en grande quantité mais à la fois la provenance et la nature de ces séquences restent indéterminées. Notamment, nos analyses montrent que l'hypothèse généralement admise que la différence marquée de composition en nucléotides des gènes transférés est due au fait qu'ils proviennent d'un hôte éloigné explique mal les caractèristiques de la plupart d'entre eux. D'autre part, pour la grande majorité des

familles protéiques dont nous disposons pour faire de la phylogénie, nous ne pouvons pas considérer que l'échantillonnage taxonomique et la conservation du signal phylogénétique sont suffisant pour résoudre la phylogénie et attribuer l'incongruence des arbres à des transferts horizontaux est une sur-interprétation des données.

L'absence de chaînons manquants fossiles dans les différentes couches géologiques a longtemps été interprétée seulement en terme de lacunes des archives paléontologiques. Après des siècles de fouilles, Gould et Eldredge (Gould et Eldredge, 1993) ont fini par proposer que ces lacunes étaient un résultat biologique : l'évolution procéderait par saut rapide entre de longues periodes de « stases » morphologiques et l'observation des stades de transition serait impossible du fait de leur faible durée à l'echelle des temps géologiques. Cette théorie des « équilibres ponctués » a eu un apport considérable à notre vision de l'évolution notamment du fait des débats qu'elle a suscités. Après à peine une décennie d'étude des génomes, de nombreux microbiologistes semblent déjà avoir tranché le débat qui aurait pu avoir eu lieu sur l'abondance et la nature des gènes transférés horizontalement : ils ont largement pris comme un résultat ce qui pourrait n'être dû qu'aux lacunes de nos données sur la biodiversité des procaryotes. De l'ordre de 5000 espèces de procaryotes ont été décrites à ce jour, mais les techniques moléculaires d'analyse de la composition des communautés microbiennes dans l'environnement suggèrent qu'elles ne représentent qu'une partie infime de la diversité réelle (Rossello-Mora et Amann, 2001). Dans ces conditions, notre capacité à reconstruire des phylogénies souffre de l'absence de ces « chaînons manquants » car un meilleur échantillonnage de la diversité permettrait probablement d'éviter un certain nombre d'artefacts méthodologiques. Cependant, puisqu'il faut faire avec ces lacunes, il est nécessaire de mettre au point des méthodes qui, non seulement permettent de prendre en compte la grande quantité de données disponibles, mais également identifient les gènes qui apportent un signal phylogénétique dans le jeu de données considéré.

Perspectives

5 Perspectives

Les perspectives de ce travail sont nombreuses : la phylogénie des procaryotes est encore loin d'être résolue et notre compréhension des mécanismes à l'œuvre dans l'évolution des génomes, notamment *via* l'acquisition de cette classe très particulière de gènes qui constituent visiblement la majorité de l'ADN transféré dans les génomes, est encore très parcellaire. Cependant, j'espère que les résultats présentés ici montrent qu'il n'est pas vain de rechercher cette phylogénie. L'apport de nouvelles séquences, et notamment une meilleure représentation de la diversité des procaryotes devrait permettre d'améliorer la résolution des relations profondes entre les divisions bactériennes et l'approche par superarbre devrait y contribuer. Plus techniquement, le maintien d'HOBACGEN-CG et surtout l'automatisation de certaines étapes de recherche des familles de gènes utilisables pour la reconstruction (reconnaissance des familles qui contiennent des paralogies) sont un des axes de développement à privilégier. De plus, la méthode de superarbre dépend de la qualité des arbres qui lui sont fournis, et l'incorporation de méthodes plus performantes (comme par exemple le maximum de vraisemblance prenant en compte l'hétérogénéité des taux d'évolution entre sites) à la procédure de reconstruction devrait améliorer grandement les résultats. Une méthode alternative, abondamment abordée dans cette thèse, est celle de la concaténation des séquences. Celle-ci nécessite d'une part d'identifier les familles portant des informations congruentes, et le test d'ILD-BIONJ sous réserve d'amélioration du calcul des distances pourrait y contribuer. Une méthode alternative qui semble particulièrement performante a été proposée par Brochier *et al.* (Brochier, *et al.*, 2002) et Matte-Tailliez *et al.* (Matte-Tailliez, *et al.*, 2002). Mais quelle que soit la méthode de sélection des gènes à concaténer, les méthodes de phylogénie ont besoin d'être adaptées pour pouvoir prendre en compte la diversité des modes d'évolution des gènes composants ces super-alignements.

L'analyse intrinsèque des génomes révèle une régionalisation du chromosome chez certaines espèces, et notamment l'importance de la localisation des gènes dans le génome sur la manière dont ils évoluent. Les hypothèses que nous avons émises peuvent être testées de diverses manières. D'abord, comme nous l'avons déjà suggéré, par l'analyse par parcimonie de génomes complets proches. Un nombre conséquent de gènomes de diverses souches d'*E. coli* devrait être disponible à court terme et permettre de vérifier s'il existe un taux de

mutation plus fort et un biais mutationnel vers A+T dans la région de terminus de réplication chez cette espèce. D'autre part, le mécanisme moléculaire dont nous avons suggéré qu'il pourrait être à l'origine de cette régionalisation peut être testé, du point de vue bioinformatique, en essayant de mettre en relation la présence d'un mécanisme de type *ter/*Tus et la structuration du G+C observée chez diverses espèces.

Enfin il reste à comprendre les caractéristiques surprenantes des gènes horizontalement transférés. Quelle est leur nature, leur fonction, leur provenance ? Quel est le mécanisme qui les maintient dans les populations, une pression de sélection ou simplement un fort taux d'insertion ? Sont-ils seulement exprimés ? Est-ce que le fait qu'ils soient fréquemment orphelins est seulement dû à notre perception biaisée de la biodiversité ? De nouveau, la disponibilité de nombreux génomes complets de souches d'une même espèce, la recherche de ces gènes dans d'autres génomes non encore séquencés et leur analyse fonctionnelle devraient nous permettre de comprendre leur dynamique et leur représentation phylogénétique.

ANNEXE A : Mécanismes d’échanges d’ADN chez les bactéries

ANNEXE A : Mécanismes d'échanges d'ADN chez les bactéries

Il est intéressant de préciser quelques détails des mécanismes d'échanges d'ADN entre les bactéries, en s'attachant plus particulièrement à la chaîne d'événements qui conduisent à l'insertion dans le chromosome de gènes autres que les gènes de plasmides, de virus ou de transposons. Comme détaillé dans le chapitre 1, les trois grands types de systèmes permettant aux bactéries d'intégrer de l'ADN étranger étaient connus dès les années 1950. Il s'agit de la conjugaison, qui est à première vue le mécanisme le plus analogue au sexe des eucaryotes en ce sens qu'il nécessite un contact entre deux individus ; de la transformation qui consiste en l'internement d'une molécule d'ADN libre du milieu ; et de la transduction qui est due à l'intervention d'un troisième protagoniste : une particule virale véhiculant l'ADN transformant.

La conjugaison

Certains plasmides comme le plasmide F d'*Escherichia coli* assurent leur transmission horizontale en apportant à la cellule la capacité à conjuguer avec une autre. Le plasmide F code près de 100 gènes dont 20 (les gènes *tra)* participent à la fonction de transfert du plasmide. Les bactéries porteuses du plasmide sont dites F+ et sont capables de former un pilus sexuel, qui peut se fixer à une cellule dépourvue de plasmide (F-). Ce pilus se raccourcit alors pour rapprocher les deux cellules et le plasmide F se réplique en envoyant sa copie dans la bactérie F-. Ce transfert se fait grâce à l'existence d'une séquence *oriT* portée par le plasmide, et à certaines enzymes formant le « relaxosome » capable de reconnaître spécifiquement cette séquence. La bactérie F- devient donc elle aussi F+. Dans cette opération, seuls les gènes portés par le plasmide (mais pas seulement les gènes *tra*) sont concernés par le transfert. D'autres plasmides non conjugatifs portant la séquence *oriT* et codant le relaxosome approprié peuvent également être recrutés ainsi par un plasmide conjugatif tel F. Mais parfois, le plasmide F s'insère dans le chromosome bactérien. Cette intégration peut se faire par différents mécanismes, à différents sites dans le chromosome. Les bactéries porteuses d'un plasmide conjugatif intégré dans le génome sont dites Hfr (pour haute fréquence de recombinaison). La conjugaison implique alors également le chromosome bactérien. Le transfert commence par la séquence *oriT*, et entraîne les parties du chromosome

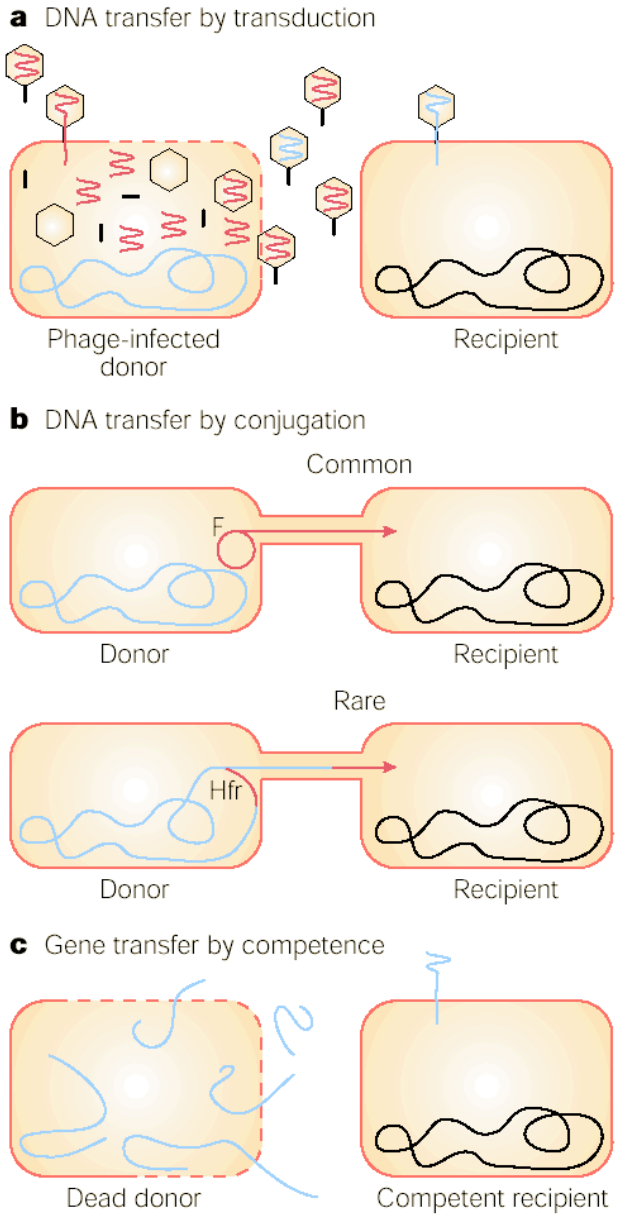


Fig. A1 : Le transfert d'ADN par transduction (a), conjugaison (b) et transformation (c). Extrait de Redfield, 2001.

bactérien adjacentes à cette séquence. La partie transférée peut être plus ou moins importante selon le temps de contact entre les cellules. La bactérie receveuse n'acquerra pas la capacité d'induire une conjugaison avec une autre bactérie car le plasmide n'est pas transmis dans sa totalité lors de cette opération. Cependant, l'ADN exogène peut recombiner avec le chromosome et des gènes peuvent ainsi être transférés (voir fig. A1). Le mécanisme de conjugaison semble pouvoir se faire entre individus d'espèces très éloignées. Il est tellement peu spécifique qu'il a été observé entre une bactérie et un eucaryote : une *E. coli* Hfr est capable de transférer son ADN à *Saccharomyces cerevisiae* ! L'étape limitante dans ce cas étant le processus de recombinaison homologue entre la séquence transmise et les chromosomes de la levure.

Les éléments transposables bactériens (ou IS pour « Insertion Sequence ») peuvent profiter de plasmides conjugatifs pour se multiplier dans de nouveaux génomes. Il arrive fréquemment qu'ils forment des éléments composites en emportant des gènes du chromosome bactérien. Ils peuvent alors transposer dans un plasmide et ainsi disséminer d'autres gènes que les leurs. Certains éléments transposables peuvent également porter les gènes induisant la conjugaison. On les trouve notamment chez les bactéries Gram-négative. C'est le cas par exemple du transposon *Tn916*, qui s'insère dans le génome bactérien, mais peut également se circulariser en emportant avec lui des séquences flanquantes du chromosome. Possédant la séquence *oriT* et les gènes

nécessaires pour induire la conjugaison, il peut se transmettre à une autre cellule de manière analogue au plasmide F, emportant avec lui des gènes de son hôte précédent.

Les plasmides, quel que soit le mode de transmission, sont très fréquemment responsables du transfert de gènes de résistance à des antibiotiques. Cela constitue pour eux, comme chaque biologiste moléculaire pratiquant le clonage de gènes le sait, un mécanisme de maintien dans l'hôte particulièrement efficace dans certaines conditions.

La transduction

Les bactériophages sont également un moyen efficace de transfert d'ADN. Leur cycle se déroule généralement comme suit : après que la particule virale ait injecté le matériel génétique dans la cellule, soit ce génome est intégré de manière réversible dans le chromosome bactérien et y persiste plus ou moins longtemps, on dit alors qu'il est sous forme de prophage, soit il est exprimé et répliqué : la cellule n'est alors plus dévouée qu'à la fabrication de particules virales et à la réplication du génome du phage. L'ADN de cette cellule est fragmenté et elle finit par être lysée, libérant ainsi de nouvelles particules infectieuses. Lors de l'encapsidation, il peut arriver que de l'ADN fragmenté du génome bactérien soit incorporé en place du génome viral. Dans ce cas, la particule infectieuse pourra injecter dans une nouvelle cellule tout autre chose que l'ADN du phage. On appelle ce mécanisme transduction généralisée. Les génomes de phage ont généralement des tailles qui vont de quelques kilobases à quelques dizaines de kilobases, ce qui peut permettre le transfert de quelques dizaines de gènes.

Certains phages ont une grande spécificité d'hôte, et même de site d'insertion pour leur forme prophage. C'est le cas notamment du phage λ d'*E. coli*, qui possède dans son génome un site *attP* homologue du site *attB* sur le génome bactérien et qui permet son intégration par recombinaison homologue. Son excision du chromosome bactérien se fait la plupart du temps en reconstituant parfaitement les deux sites *attP* et *attB*. Cependant, il arrive qu'avec l'ADN du phage soient encapsidés les gènes *gal* (impliqué dans le métabolisme du galactose) et *bio* (impliqué dans la synthèse de la biotine) flanquant le site *attB*. Ainsi, même la transduction spécialisée peut entraîner le transfert de gènes non phagiques. A l'inverse, certains bactériophages comme le phage Mu possèdent un large spectre d'hôte et peuvent s'insérer dans n'importe quel site du génome. Il semble cependant que les gènes d'ARN de

transfert soient des sites privilégiés d'insertion pour ces phages, bien que la raison en soit encore obscure. Le bactériophage Mu possède un mécanisme d'insertion très proche de certains transposons, qui lui permet non seulement de s'insérer, mais également de transposer sous sa forme prophage. Lors de son excision, le génome du phage emporte un peu des régions flanquant son site d'insertion, qui peuvent se retrouver encapsidés et injectés dans le prochain hôte. Contrairement au cas du phage λ, chacun des gènes d'un génome peut donc être entraînés avec le génome du phage dans la transduction spécifique de Mu.

Les génomes de bactériophages contiennent fréquemment des gènes conférant à la bactérie hôte un grand avantage dans certains milieux, notamment des gènes de pathogénicité regroupés en « îlots ». Les souches virulentes de *Corynebacterium diphteria*, responsables de la diphtérie, ne diffèrent des souches bénignes que par la présence d'un bactériophage appelé corynephage (β ou ω). Ce bactériophage porte des gènes capables de fabriquer une toxine qui provoque la destruction des cellules du sujet abritant *C. diphteria*, ce qui provoque un apport important de nutriments (et notamment de fer) à la bactérie. De nombreux autres exemples de bactéries pathogènes de l'homme doivent leur virulence à des bactériophages : c'est le cas notamment de *Streptococcus pyogenes* et du bactériophage T12, responsables à eux deux de la scarlatine ; de *Vibrio cholerae* et de ses phages CTXΦ et VPIΦ. La présence d'îlots de pathogénicité souvent à proximité de gènes d'ARN de transfert (ARNt) chez de nombreuses bactéries comme la souche uropathogénique d'*E.coli* UPEC, suggère que leur dissémination pourrait être principalement assurée par les phages, dont les ARNt sont des sites privilégiés d'action des intégrases.

La transformation

La transformation est le mécanisme par lequel de l'ADN libre peut entrer dans la cellule. Une cellule capable d'intégrer de l'ADN par ce biais est dite compétente. La compétence peut être induite artificiellement chez la plupart des bactéries, et même chez des eucaryotes par des traitements spéciaux comme l'électroporation ou le traitement au $CaCl_2$, mais certaines bactéries comme *Streptococcus pneumoniae, Neisseria gonorrhoeae* et *Haemophilus influenzae* passent spontanément à l'état de compétence dans certaines conditions. La transformation chez les bactéries est un mécanisme actif qui utilise des gènes spécifiques (notamment les gènes *com*). De très nombreuses bactéries dont on n'a jamais

observé de compétence naturelle possèdent ces gènes, ce qui suggère qu'elles peuvent entrer naturellement en compétence dans des conditions encore inconnues. Le mécanisme diffère d'une bactérie à l'autre : *Haemophilus influenzae* par exemple reconnaît une séquence spécifique de neuf paires de bases à sa surface, et intègre ensuite l'ADN sous forme double brin. Cependant, certaines bactéries ne semblent pas posséder de tel mécanisme de reconnaissance de séquences et internalisent l'ADN en dégradant l'un des brins. Dans ce cas, pour un ADN non auto-réplicatif, le succès du transfert dépendra du degré de similarité de la séquence avec le chromosome. En effet, dans ce mode de transfert horizontal particulièrement (mais pas seulement), l'étape de recombinaison homologue est critique. D'ordinaire, seul un ADN présentant un fort degré de similarité avec le chromosome sur une portion de séquence de longueur variable selon les bactéries, pourra recombiner. Cependant, dans certaines conditions, comme pendant un stress important ou lorsque des gènes contrôlant la spécificité de l'appariement des deux brins d'ADN sont mutés, des événements de recombinaison hétérologue peuvent avoir lieu, et provoquer l'intégration de l'ADN d'une espèces éloignée.

ANNEXE B : Brefs rappels de phylogénie moléculaire

ANNEXE B : Brefs rappels de phylogénie moléculaire

Tous les organismes vivants possèdent un ancêtre commun. Ceci implique que les ressemblances que l'on observe entre eux sont l'indice de leur proximité phylogénétique. Historiquement basés sur les caractères morphologiques, les concepts de la phylogénie ont été adaptés aux séquences des macromolécules informatives (protéines et ADN), ceci depuis les travaux de Zuckerkandl et Pauling, 1965. Ainsi, le concept d'homologie et le principe des connexions peuvent s'appliquer en biologie moléculaire pour identifier les traces des évènements évolutifs dans les gènes et ainsi retracer leur phylogénie. On peut faire l'hypothèse de la correspondance entre phylogénie des gènes et phylogénie des espèces sous certaines conditions que nous allons voir. Les avantages de la phylogénie moléculaire sur la morphologie sont multiples : elle permet notamment d'inférer l'histoire d'espèces pour lesquelles peu de caractères morphologiques sont disponibles et dont les caractères phénotypiques montrent un degré de convergence important (typiquement, les microorganismes, voir section 1.3). En outre, elle donne accès à une quantité très supérieure de caractères exploitables pour la phylogénie. Cependant, certains problèmes spécifiques au matériel moléculaire se posent et ont nécessité l'invention d'un certain nombre de concepts. Nous allons détailler les plus importants d'entre eux et voir brièvement certaines limites des méthodes de phylogénie moléculaire.

Qu'est-ce qu'un arbre ?

Les phénomènes d'individualisation des espèces (spéciation) et des gènes (spéciation et duplication) peuvent se représenter sous la forme d'arbres binaires, constitués de nœuds et de branches. En phylogénie moléculaire, les nœuds terminaux, ou feuilles représentent les séquences actuelles, les nœuds internes représentent les séquences hypothétiques. Les branches représentent les relations d'ascendance ou de descendance entre ces entités, et peuvent êtres internes (si elles lient deux nœuds internes) ou terminales (si elles lient un nœud interne à un nœud terminal). On utilise le terme de topologie pour désigner la structure de

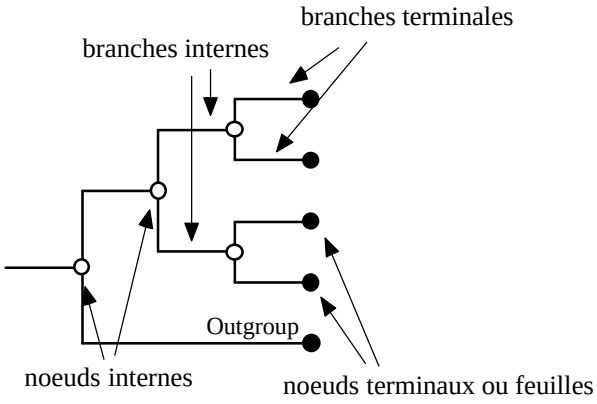


Fig. B1 : un arbre phylogénétique type. L'Outgroup ou groupe externe permet de raciner l'arbre et d'orienter les évènements dans l'arbre.

l'arbre, c'est-à-dire l'ordre de branchement des séquences qui le composent. Les branches de l'arbre sont caractérisées par les longueurs qui représentent la quantité de changements évolutifs inférés sur ses branches (généralement exprimés en nombre de substitutions par site). Pour pouvoir interpréter un arbre phylogénétique, il est absolument nécessaire de le raciner afin de pouvoir orienter dans le temps la suite des évènements qu'il décrit. Un outgroup ou groupe externe (groupe de séquences dont la divergence est antérieure à la radiation du groupe étudié) est généralement utilisé pour positionner la racine. Une fois orientés, les arbres décrivent des groupes monophylétiques (groupes dont toutes les séquences sont plus proches entre elles qu'elles ne le sont de n'importe quelle séquence extérieure à ce groupe)(fig. B2).

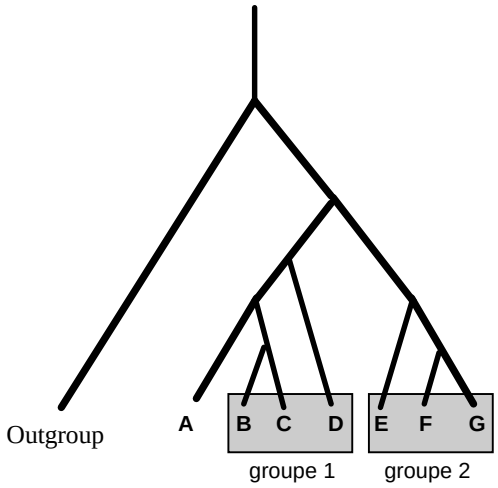


Fig. B2 :monophylie et paraphylie : Le groupe 2 est monophylétique. Par contre, le groupe 1 ne l'est pas car les séquences B et C sont plus proches de A qu'elles ne le sont de D. On dit que ce groupe est paraphylétique.

L'homologie

Deux séquences sont homologues si elles possèdent une séquence ancestrale commune. Cependant, les gènes peuvent acquérir leur indépendance évolutive de deux manières : par spéciation et par duplication. Il existe en effet des gènes qui sont en plusieurs copies dans les génomes et dont il est possible de retracer l'histoire en phylogénie moléculaire. Ainsi, en phylogénie moléculaire, un noeuds peut représenter un évènement de spéciation ou de duplication. Cette particularité a nécessité la définition de deux types de relations d'homologie : l'orthologie et la paralogie. Deux gènes qui ont acquis leur indépendance évolutive à la suite d'un évènement de spéciation sont orthologues. Ceux qui ont acquis cette indépendance à la suite d'une duplication sont paralogues. Ce concept est important car on voit bien que seuls des orthologues peuvent décrire l'histoire des espèces. La confusion entre des gènes paralogues et

orthologues peut avoir des conséquences importantes sur la phylogénie notamment si les duplication sont anciennes. L'identification des paralogies dans un arbre, si l'on veut retracer l'histoire des espèces, est donc primordiale mais parfois malaisée du fait que certains gènes peuvent avoir été perdus ou ne pas avoir été séquencés (Fig B3).

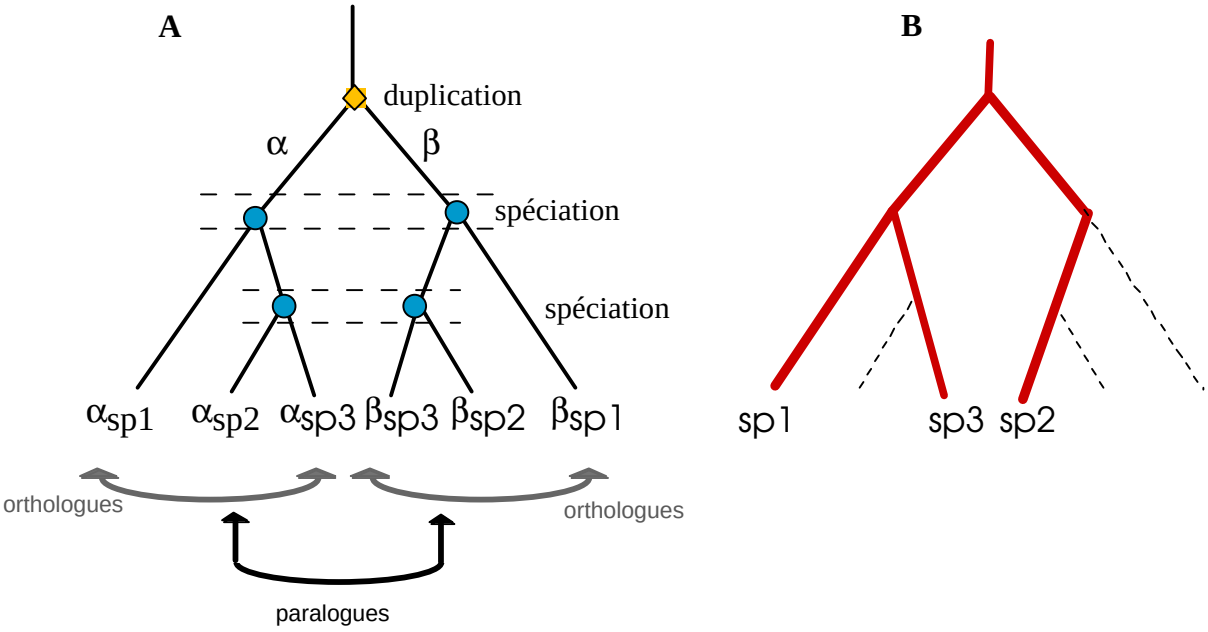


Fig. B3 : Orthologie et Paralogie. **A** : Un gène a subi une duplication chez l'ancêtre commun à trois espèces actuelles (sp1, sp2 et sp3). Les gènes α sont orthologues entre eux (les noeuds les plus récents qui relient chacun de ces gènes deux a deux sont tous des noeuds de spéciation) De même pour les gènes β. Les gènes α et β sont paralogues (les noeuds les plus récents qui relient chacun de ces gènes deux a deux sont tous des noeuds de duplication). **B** : Si certains de ces gènes manquent (soit parce que les données sont lacunaires, soit du fait de pertes de certains gènes), il devient impossible de différencier noeud de spéciation et de duplication et la phylogénie des espèces inférée est fausse.

Information et saturation

Au cours des temps évolutifs, les séquences accumulent indépendamment des différences (« substitutions » de bases dans l'ADN ou d'acides aminés dans les protéines). Ce sont sur ces différences que s'appuie la reconstruction phylogénétique. Par exemple dans la fig. B4, pour le site 1, l'absence de substitutions ne permet de résoudre aucune des branches de l'arbre. Au site 2, au contraire, la substitution qui s'est produite pourra facilement être utilisée pour reconstruire l'arbre et plaidera pour le regroupement des espèces 1 et 2. Ce type de site contient une information phylogénétique que s'attachent à exploiter les différentes méthodes de reconstruction. Par contre, le site 3 a subi un nombre important de substitutions

au cours de son histoire et il ne contient plus d'information. On appelle ce phénomène saturation. Dans ce cas, aucune espèce ne partage d'état de sites en commun, cependant la saturation peut produire des convergences entre sites, notamment dans les séquences d'ADN où il n'existe que quatre états possibles pour un site (A, T, C et G).

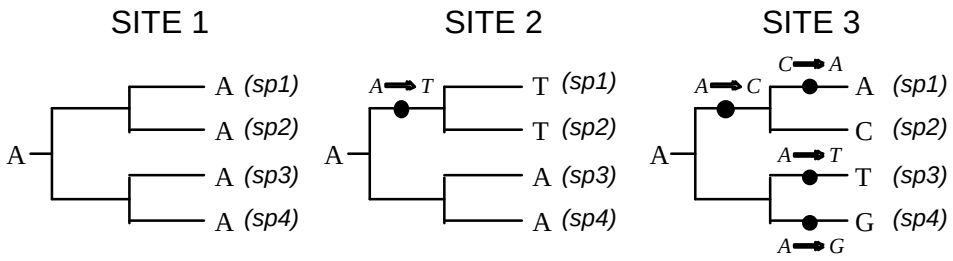


Fig. B4 : Trois sites homologues dans quatre espèces. Seul le site 2 contient une information exploitable pour la phylogénie. Les substitutions multiples sur le site 3 conduisent à une perte du signal phylogénétique (saturation).

La saturation du signal est un problème qui se pose particulièrement à des échelles évolutives importantes comme pour résoudre la phylogénie des bactéries. Certains gènes relativement peu contraints apportent une résolution à des faibles distances évolutives mais ne permettent pas d'inférer des phylogénies plus anciennes. A l'inverse, les gènes les plus conservés (comme par exemple l'ARN ribosomal) peuvent apporter une information sur les liens de parenté entre organismes éloignés, mais plus difficilement entre espèces proches. A l'échelle de la phylogénie du vivant, rares sont les gènes pour lesquels le signal phylogénétique n'est pas saturé.

Le problème majeur lié au phénomène de saturation est qu'il peut se produire de manière plus ou moins intense entre les lignées évolutives. Le cas le plus dramatique est connu sous le nom de phénomène d'attraction des longues branches (LBA pour « Long Branch Attraction »). Il a été décrit dès 1978 par Felsenstein (Felsenstein, 1978). On peut en effet montrer de manière analytique, dans le cas simple d'un arbre à quatre taxons, que des branches ayant des taux d'évolution très supérieurs aux autres vont se retrouver artéfactuellement regroupées (fig B5). Le phénomène peut se comprendre facilement de manière intuitive : si au sein d'un groupe une lignée tend à accumuler de nombreuses substitutions dans ces gènes, ceux-ci vont finir par être tellement différents des gènes des

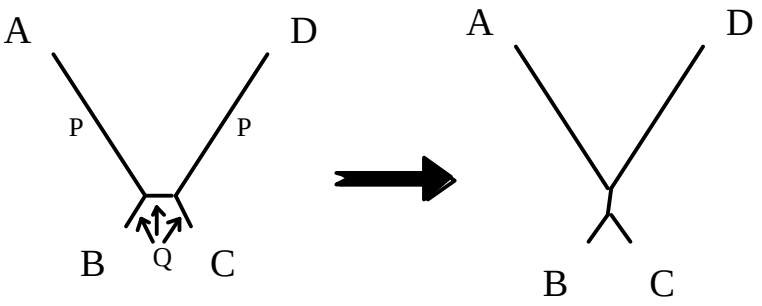


Fig. B5 : Le phénomène d'attraction des longues branches décrit par Felsenstein, 1978. Dans le cas d'un arbre vrai (à gauche) où certaines lignées ont des taux d'évolution très forts et d'autres des taux beaucoup plus faibles (P >> Q), les méthodes de phylogénétiques retrouvent invariablement l'arbre de droite, où les lignées ayant des taux d'évolution forts sont regroupées.

autres membres du groupe qu'ils vont s'en trouver exclus par les méthodes de reconstruction phylogénétique. Par exemple, dans l'arbre de la fig.B5, les sites portant une information sur la branche interne auront une forte probabilité de muter de nouveau dans les branches menant à A et D et le signal supportant l'arbre vrai sera perdu. Ce phénomène pourrait donc plus légitimement être qualifié de phénomène d'exclusion des longues branches. Lorsque les différences de taux d'évolution sont très importantes, les séquences affectées par le phénomène d'attraction des longues branches se groupent avec la séquence la plus divergente de l'arbre, c'est-à-dire le groupe externe. Plusieurs groupes ont été considérés comme primitifs car ils se branchaient à la base de l'arbre, et ont ensuite été replacés au sein de l'arbre par des analyses plus précises. C'est le cas notamment des microsporidies, d'abord considérées sur la base de phylogénies moléculaires comme des eucaryotes ayant émergé très précocement, et dont on sait aujourd'hui qu'elles sont en fait un groupe de champignons ayant des taux d'évolution extrêmes du fait de leur mode de vie parasitaire (voir par exemple Thomarat, 2002).

Méthodes de reconstruction

Il existe trois grandes classes de reconstruction phylogénétiques : la méthode de parcimonie, les méthodes de distance et les méthodes de maximum de vraissemblance. Chacune de ces méthodes permet de choisir l'arbre qui permet d'optimiser un critère. Cependant, le nombre d'arbres devenant rapidement astronomique avec le nombre d'espèces qu'ils contiennent, les méthodes utilisent des heuristiques qui, si elles ne garantissent pas de trouver le meilleur arbre, permettent de trouver des arbres proches de celui-ci pour le critère considéré.

Le critère minimisé par la méthode de parcimonie est le nombre de changements d'états de caractères (états des sites) que chaque arbre nécessite d'inférer en fonction de la matrice (l'alignement des séquences). Cette méthode est relativement lente à évaluer les arbres, même si l'on utilise une heuristique.

Les méthodes de distances (type Neighbor-Joining ou BIONJ) nécessitent de transformer la matrice de caractères (alignement) en une matrice de distance par comparaison des lignes de la matrice (séquences) deux à deux. En phylogénie moléculaire, ces comparaisons peuvent se faire sur la base d'un modèle évolutif qui décrit de manière statistique le processus évolutif agissant sur la séquence. Le modèle utilisé dans cette thèse pour la reconstruction phylogénétique est le modèle JTT (Jones, *et al.*, 1992). Il sagit d'une matrice qui décrit les probabilités de substitution d'un acide aminé par un autre basée sur l'analyse d'un nombre important d'alignements protéiques. Une fois les distances calculées, les méthodes de distance permettront de choisir l'arbre dont la somme des longueurs de branches est minimale. Ces méthodes permettent en général de calculer les arbres très rapidement.

De même que les méthodes de distances, les méthodes de maximum de vraisemblance permettent d'utiliser un modèle évolutif. Cependant, celles-ci ne comparent pas les séquences deux à deux, mais estiment la vraisemblance de chaque site pour chaque topologie au regard du modèle évolutif choisi. La topologie choisie par la méthode sera celle qui maximise la vraisemblance de l'alignement. De même, nous avons utilisé le modèle JTT (Jones, *et al.*, 1992) pour évaluer les arbres de maximum de vraisemblance. Cette méthode nécessite des temps de calculs très importants.

Support statistique des phylogénies

Toutes les méthodes de reconstruction phylogénétique, notamment les méthodes que nous avons utilisées, fournissent un arbre phylogénétique final. Il est important d'estimer quel est le support statistique de chacune des branches internes de cet arbre. L'on utilise pour cela

le plus souvent la méthode de *bootstrap*. Cette méthode consiste à simuler à partir de l'alignement de départ, un nombre important d'alignements (au moins 500) de même taille par tirage aléatoire avec remise. A partir de chacun de ces alignements simulés, un arbre est reconstruit et l'on peut reporter sur chaque branche interne de l'arbre de départ, le nombre de fois où cette branche a été retrouvée dans les données simulées. Cet indice de *bootstrap* indique donc la robustesse statistique de la branche interne. La nécessité de reconstruire plusieurs centaines d'arbres fait que la méthode de *bootstrap* est difficilement utilisable pour évaluer les arbres de maximum de vraisemblance que nous avons reconstruits au cours de cette thèse. Nous avons donc utilisé pour cette méthode un indice qui estime les valeurs de *bootstrap*, le RELL, défini par Kishino et al. (1990).

Article 1 : A phylogenomic approach to bacterial phylogeny : evidence of a core of genes sharing a common history

Daubin Vincent, Gouy Manolo et Perrière Guy

Publié dans *Genome Research* (2002) 12 : 1080-1090

Article 2 : G+C3 structuring along the genome : a common feature in prokaryotes

Daubin Vincent et Perrière Guy

Accepté dans *Molecular Biology and Evolution*

Références bibliographiques

Références bibliographiques

Achaz, G., Rocha, E. P., Netter, P. and Coissac, E. (2002). Origin and fate of repeats in bacteria. *Nucleic Acids Res* **30,** 2987-94

Achenbach-Richter, L., Gupta, R., Stetter, K. O. and Woese, C. R. (1987). Were the original eubacteria thermophiles? *Syst Appl Microbiol* **9,** 34-9

Adachi, J. and Hasegawa, M. (1996). Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships. *Mol Phylogenet Evol* **6,** 72-6

Alimi, J. P., Poirot, O., Lopez, F. and Claverie, J. M. (2000). Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from Escherichia coli K-12 MG1655. *Genome Res* **10,** 959-66

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389-402

Andersson, S. G. and Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* **54,** 198-210

Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. and Koonin, E. V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet* **14,** 442-4

Arber, W. and Kehnlein, U. (1967). Mutational loss of B-specific restriction of the bacteriophage. *fd. Path. Micro.* **30,** 946-952

Avery, O. T., Macleod, C. M. and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumonococcal types. Induction of transformation by a deoxyribo-nucleic acid fraction isolated from pnuemococcus type III. *J. Exp. Med.* **79,** 137-157. *(In Microbiology: A Centenary Perspective, edited by Wolfgang K. Joklik, ASM Press. 1999, p.116)*

Baldauf, S. L., Palmer, J. D. and Doolittle, W. F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc Natl Acad Sci U S A* **93,** 7749-54

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M. and Philippe, H. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A* **99,** 1414-9

Barns, S. M., Delwiche, C. F., Palmer, J. D. and Pace, N. R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci U S A* **93,** 9188-93

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41,** 3-10

Beadle, G. and Tatum, E. (1941). Genetic control of biochemical reactions in Neurospora. *Proc. Nat. Acad. Sci.* **27,** 499-506. *(In Microbiology: A Centenary Perspective, edited by Wolfgang K. Joklik, ASM Press. 1999, p.308)*

Bellgard, M. I., Itoh, T., Watanabe, H., Imanishi, T. and Gojobori, T. (1999). Dynamic evolution of genomes and the concept of genome space. *Ann N Y Acad Sci* **870,** 293-300

Beltran, P., Musser, J. M., Helmuth, R., Farmer, J. J., 3rd, Frerichs, W. M., Wachsmuth, I. K., Ferris, K., McWhorter, A. C., Wells, J. G., Cravioto, A. and et al. (1988). Toward a population genetic analysis of Salmonella: genetic diversity and relationships among strains of serotypes S. choleraesuis, S. derby, S. dublin, S. enteritidis, S. heidelberg, S. infantis, S. newport, and S. typhimurium. *Proc Natl Acad Sci U S A* **85,** 7753-7

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L. (2002). GenBank. *Nucleic Acids Res* **30,** 17-20

Benzécri, J. (1973). *L'analyse de données*.

Berg, O. G. and Kurland, C. G. (1997). Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270,** 544-50

Berka, R. M., Hahn, J., Albano, M., Draskovic, I., Persuh, M., Cui, X., Sloma, A., Widner, W. and Dubnau, D. (2002). Microarray analysis of the Bacillus subtilis K-state: genome-wide expression changes dependent on ComK. *Mol Microbiol* **43,** 1331-45

Bertolla, F., Van Gijsegem, F., Nesme, X. and Simonet, P. (1997). Conditions for natural transformation of Ralstonia solanacearum. *Appl Environ Microbiol* **63,** 4965-8

Bierne, H., Ehrlich, S. D. and Michel, B. (1997). Deletions at stalled replication forks occur by two different pathways. *Embo J* **16,** 3332-40

Bierne, H. and Michel, B. (1994). When replication forks stop. *Mol Microbiol* **13,** 17-23

Bowler, L. D., Zhang, Q. Y., Riou, J. Y. and Spratt, B. G. (1994). Interspecies recombination between the penA genes of Neisseria meningitidis and commensal Neisseria species during the emergence of penicillin resistance in N. meningitidis: natural events and laboratory simulation. *J Bacteriol* **176,** 333-7

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol* **59,** 27-40

Brochier, C., Bapteste, E., Moreira, D. and Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18,** 1-5

Brochier, C. and Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417,** 244

Brochier, C., Philippe, H. and Moreira, D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* **16,** 529-33

Brown, E. W., Kotewicz, M. L. and Cebula, T. A. (2002). Detection of recombination among Salmonella enterica strains using the incongruence length difference test. *Mol Phylogenet Evol* **24,** 102-20

Brown, J. R. and Doolittle, W. F. (1997). Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61,** 456-502

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. and Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28,** 281-5

Brumbley, S. M., Carney, B. F. and Denny, T. P. (1993). Phenotype conversion in Pseudomonas solanacearum due to spontaneous inactivation of PhcA, a putative LysR transcriptional regulator. *J Bacteriol* **175,** 5477-87

Bryant, D. and Steel, M. (2001). Constructing optimal trees from quartets. *Journal of Algorithms* **38,** 237-259

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature* **325,** 728-30

Bussiere, D. E. and Bastia, D. (1999). Termination of DNA replication of bacterial and plasmid chromosomes. *Mol Microbiol* **31,** 1611-8

Capiaux, H., Cornet, F., Corre, J., Guijo, M. I., Perals, K., Rebollo, J. E. and Louarn, J. M. (2001). Polarization of the Escherichia coli chromosome. A view from the terminus. *Biochimie* **83,** 161-70

Capiaux, H., Lesterlin, C., Perals, K., Louarn, J. M. and Cornet, F. (2002). A dual role for the FtsK protein in Escherichia coli chromosome segregation. *EMBO Rep* **3,** 532-6

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17,** 540-52

Cavalier-Smith, T. (1987). The origin of eukaryotic and archaebacterial cells. *Ann N Y Acad Sci* **503,** 17-54

Clarke, G. D., Beiko, R. G., Ragan, M. A. and Charlebois, R. L. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* **184,** 2072-80

Confalonieri, F., Elie, C., Nadal, M., de La Tour, C., Forterre, P. and Duguet, M. (1993). Reverse gyrase: a helicase-like domain and a type I topoisomerase in the same polypeptide. *Proc Natl Acad Sci U S A* **90,** 4753-7

Cox, M. M. (2001). Recombinational DNA repair of damaged replication forks in Escherichia coli: questions. *Annu Rev Genet* **35,** 53-82

Cox, M. M., Goodman, M. F., Kreuzer, K. N., Sherratt, D. J., Sandler, S. J. and Marians, K. J. (2000). The importance of repairing stalled replication forks. *Nature* **404,** 37-41

Cunningham, C. W. (1997a). Can three incongruence tests predict when data should be combined? *Mol Biol Evol* **14,** 733-40

Cunningham, C. W. (1997b). Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst Biol* **46,** 464-78

Darlu, P. and Lecointre, G. (2002). When does the incongruence length difference test fail? *Mol Biol Evol* **19,** 432-7

Daubin, V., Gouy, M. and Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12,** 1080-90

Daubin, V., Gouy, M. and Perrière, G. (2001). Bacterial molecular phylogeny using supertree approach. *Genome Inform Ser Workshop Genome Inform* **12,** 155-64

Dayhoff, M. O. and Schwartz, R. M. (1981). Evidence on the origin of eukaryotic mitochondria from protein and nucleic acid sequences. *Ann N Y Acad Sci* **361,** 92-104

de la Cruz, F. and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* **8,** 128-33

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J. and Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature* **392,** 353-8

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., Radman, M. and Matic, I. (2000). Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103,** 711-21

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,** 1350-3

Dhavan, G. M., Crothers, D. M., Chance, M. R. and Brenowitz, M. (2002). Concerted binding and bending of DNA by Escherichia coli integration host factor. *J Mol Biol* **315,** 1027-37

Dolphin, K., Belshaw, R., Orme, C. D. and Quicke, D. L. (2000). Noise and incongruence: interpreting results of the incongruence length difference test. *Mol Phylogenet Evol* **17,** 401-6
Doolittle, W. F. (1999a). Lateral genomics. *Trends Cell Biol* **9,** M5-8
Doolittle, W. F. (1999b). Phylogenetic classification and the universal tree. *Science* **284,** 2124-9
Dowson, C. G., Coffey, T. J., Kell, C. and Whiley, R. A. (1993). Evolution of penicillin resistance in Streptococcus pneumoniae; the role of Streptococcus mitis in the formation of a low affinity PBP2B in S. pneumoniae. *Mol Microbiol* **9,** 635-43
Dowton, M. and Austin, A. D. (2002). Increased congruence does not necessarily indicate increased phylogenetic accuracy--the behavior of the incongruence length difference test in mixed-model analyses. *Syst Biol* **51,** 19-31
Echenique, J. R., Chapuy-Regaud, S. and Trombe, M. C. (2000). Competence regulation by oxygen in Streptococcus pneumoniae: involvement of ciaRH and comCDE. *Mol Microbiol* **36,** 688-96
Eisen, J. A., Heidelberg, J. F., White, O. and Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1,** RESEARCH0011
Eyre-Walker, A. and Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* **21,** 4599-603
Farris, J. S., Källersjö, M., Kluge, A. G. and Bult, C. (1994). Testing significance of congruence. *Cladistics* **10,** 315-319
Felsenstein, J. (1978). Cases in which parsimony and compatibility methods wil be positively misleading. *Syst Zool* **27,** 401-10
Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5,** 164-6
Finkel, S. E. and Johnson, R. C. (1992). The Fis protein: it's not just for DNA inversion anymore. *Mol Microbiol* **6,** 3257-65
Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19,** 99-113
Fitz-Gibbon, S. T. and House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res* **27,** 4218-22
Forterre, P. (1995). Thermoreduction, a hypothesis for the origin of prokaryotes. *C R Acad Sci III* **318,** 415-22
Forterre, P. (2002). A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet* **18,** 236-7
Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C. and Labedan, B. (1992). The nature of the last universal ancestor and the root of the tree of life, still open questions. *Biosystems* **28,** 15-32
Forterre, P., Bouthier De La Tour, C., Philippe, H. and Duguet, M. (2000). Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. *Trends Genet* **16,** 152-4
Fox, G. E., Pechman, K. R. and Woese, C. R. (1977). Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int. J. Syst. Bacteriol.* **27,** 44-57. *(In In Microbiology: A Centenary Perspective, edited by Wolfgang K. Joklik, ASM Press. 1999, p.264)*
Francino, M. P. and Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends Genet* **13,** 240-5
Frank, A. C. and Lobry, J. R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238,** 65-77

Frank, A. C. and Lobry, J. R. (2000). Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16,** 560-1

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J. and Venter, J. C. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. *Nature* **390,** 580-6

Galtier, N. and Gouy, M. (1994). Molecular phylogeny of Eubacteria: a new multiple tree analysis method applied to 15 sequence data sets questions the monophyly of gram-positive bacteria. *Res Microbiol* **145,** 531-41

Galtier, N. and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol* **44,** 632-6

Galtier, N., Tourasse, N. and Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science* **283,** 220-1

Garcia-Vallve, S., Romeu, A. and Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10,** 1719-25

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14,** 685-95

Glansdorff, N. (2000). About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. *Mol Microbiol* **38,** 177-85

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T. and et al. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A* **86,** 6661-5

Golding, G. B. and Gupta, R. S. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol Biol Evol* **12,** 1-6

Goncalves, I., Robinson, M., Perriere, G. and Mouchiroud, D. (1999). JaDis: computing distances between nucleic acid sequences. *Bioinformatics* **15,** 424-5

Gould, S. J. and Eldredge, N. (1993). Punctuated equilibrium comes of age. *Nature* **366,** 223-7

Gouy, M. and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10,** 7055-74

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. (1985). ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput Appl Biosci* **1,** 167-72

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325-328

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,** r43-74

Gregg, A. V., McGlynn, P., Jaktaji, R. P. and Lloyd, R. G. (2002). Direct rescue of stalled DNA replication forks via the combined action of PriA and RecG helicase activities. *Mol Cell* **9,** 241-51

Griffith, F. (1928). The significance of pneumococcal types. *J. Hyg.* **27,** 113-159

Grishin, N. V., Wolf, Y. I. and Koonin, E. V. (2000). From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res* **10,** 991-1000

Groisman, E. A., Saier, M. H., Jr. and Ochman, H. (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the Salmonella genome. *Embo J* **11,** 1309-16

Groisman, E. A., Sturmoski, M. A., Solomon, F. R., Lin, R. and Ochman, H. (1993). Molecular, functional, and evolutionary analysis of sequences specific to Salmonella. *Proc Natl Acad Sci U S A* **90,** 1033-7

Guindon, S. and Perriere, G. (2001). Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol Biol Evol* **18,** 1838-40

Guiseppi, A., Aymeric, J. L., Cami, B., Barras, F. and Creuzet, N. (1991). Sequence analysis of the cellulase-encoding celY gene of Erwinia chrysanthemi: a possible case of interspecies gene transfer. *Gene* **106,** 109-14

Gupta, R. S. (1998a). Life's third domain (Archaea): an established fact or an endangered paradigm? *Theor Popul Biol* **54,** 91-104

Gupta, R. S. (1998b). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62,** 1435-91

Gupta, R. S. and Golding, G. B. (1993). Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J Mol Evol* **37,** 573-82

Hacker, J. and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54,** 641-79

Hahn, J., Luttinger, A. and Dubnau, D. (1996). Regulatory inputs for the synthesis of ComK, the competence transcription factor of Bacillus subtilis. *Mol Microbiol* **21,** 763-75

Hansmann, S. and Martin, W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* **50 Pt 4,** 1655-63

Hayes, W. (1952). Recombination in Bact.coli. K-12: unidirectional transfer of genetic material. *Nature* **169,** 118-119

Hershey, A. D. and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol* **36,** 39-56. *(In Microbiology: A Centenary Perspective, edited by Wolfgang K. Joklik, ASM Press. 1999, p.474)*

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266,** 383-402

Hill, T. M. (1992). Arrest of bacterial DNA replication. *Annu Rev Microbiol* **46,** 603-33

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. and Herrmann, R. (1997). Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. *Nucleic Acids Res* **25,** 701-12

Horiuchi, T. and Fujimura, Y. (1995). Recombinational rescue of the stalled DNA replication fork: a model based on analysis of an Escherichia coli strain with a chromosome region difficult to replicate. *J Bacteriol* **177,** 783-91

House, C. H. and Fitz-Gibbon, S. T. (2002). Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Mol Evol* **54,** 539-47

Huang, Y. P. and Ito, J. (1999). DNA polymerase C of the thermophilic bacterium Thermus aquaticus: classification and phylogenetic analysis of the family C DNA polymerases. *J Mol Evol* **48,** 756-69

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C. and Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417,** 63-7

Hudson, R. E., Bergthorsson, U., Roth, J. R. and Ochman, H. (2002). Effect of chromosome location on bacterial mutation rates. *Mol Biol Evol* **19,** 85-92

Huynen, M. A. and Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A* **95,** 5849-56

Huynen, M. A., Snel, B. and Bork, P. (1999). Lateral Gene Transfer, Genome Surveys, and the Phylogeny of Prokaryotes. *Science* **286,** 1443a

Ide, H., Murayama, H., Sakamoto, S., Makino, K., Honda, K., Nakamuta, H., Sasaki, M. and Sugimoto, N. (1995). On the mechanism of preferential incorporation of dAMP at abasic sites in translesional DNA synthesis. Role of proofreading activity of DNA polymerase and thermodynamic characterization of model template-primers containing an abasic site. *Nucleic Acids Res* **23,** 123-9

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151,** 389-409

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. and Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A* **86,** 9355-9

Jain, R., Rivera, M. C. and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96,** 3801-6

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8,** 275-82

Kakinuma, Y., Igarashi, K., Konishi, K. and Yamato, I. (1991). Primary structure of the alpha-subunit of vacuolar-type Na(+)-ATPase in Enterococcus hirae. Amplification of a 1000-bp fragment by polymerase chain reaction. *FEBS Lett* **292,** 64-8

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238,** 143-55

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M. and Tabata, S. (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. *DNA Res* **7,** 331-8

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1,** 598-610

Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9,** 335-43

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11,** 283-90

Karlin, S., Campbell, A. M. and Mrazek, J. (1998). Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32,** 185-225

Karlin, S. and Mrazek, J. (1997). Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A* **94,** 10227-32

Karlin, S., Mrazek, J. and Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179,** 3899-913

Kerr, A. R., Peden, J. F. and Sharp, P. M. (1997). Systematic base composition variation around the genome of Mycoplasma genitalium, but not Mycoplasma pneumoniae. *Mol Microbiol* **25,** 1177-9

Kishino, H. T., Miayata, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplast. *J Mol Evol* **31,** 151-60
Klenk, H. P., Meier, T. D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P. P. and Zillig, W. (1999). RNA polymerase of Aquifex pyrophilus: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. *J Mol Evol* **48,** 528-41
Koonin, E. V., Makarova, K. S. and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55,** 709-42
Koski, L. B. and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52,** 540-2
Kreil, D. P. and Ouzounis, C. A. (2001). Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* **29,** 1608-15
Kroll, J. S. and Moxon, E. R. (1990). Capsulation in distantly related strains of Haemophilus influenzae type b: genetic drift and gene transfer at the capsulation locus. *J Bacteriol* **172,** 1374-9
Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Danchin, A. and et al. (1997). The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390,** 249-56
Kurland, C. G. (2000). Something for everyone. Horizontal gene transfer in evolution. *EMBO Rep* **1,** 92-5
Kuzminov, A. (1995). Collapse and repair of replication forks in Escherichia coli. *Mol Microbiol* **16,** 373-84
Kyrpides, N. C. and Olsen, G. J. (1999). Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry? *Trends Genet* **15,** 298-9
Lake, J. A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331,** 184-6
Lake, J. A. and Rivera, M. C. (1994). Was the nucleus the first endosymbiont? *Proc Natl Acad Sci U S A* **91,** 2880-1
Lawrence, J. G. and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44,** 383-97
Lawrence, J. G. and Ochman, H. (1998). Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A* **95,** 9413-7
Lawrence, J. G. and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143,** 1843-60
Lecointre, G., Philippe, H., Van Le, H. L. and Le Guyader, H. (1993). Species sampling has a major impact on phylogenetic inference. *Mol Phylogenet Evol* **2,** 205-24
Lecointre, G., Rachdi, L., Darlu, P. and Denamur, E. (1998). Escherichia coli molecular phylogeny using the incongruence length difference test. *Mol Biol Evol* **15,** 1685-95
Lederberg, J. (1952). Cell genetics and hereditary symbiosis. *Physiol. Rev.* **32,** 403-430
Lederberg, J. and Tatum, E. L. (1946). Gene recombination in Eschericia coli. *Nature* **58,** 558
Lemon, K. P., Kurtser, I. and Grossman, A. D. (2001). Effects of replication termination mutants on chromosome partitioning in Bacillus subtilis. *Proc Natl Acad Sci U S A* **98,** 212-7
Levin, B. R. and Bergstrom, C. T. (2000). Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proc Natl Acad Sci U S A* **97,** 6981-5
Lewis, P. J. (2001). Bacterial chromosome segregation. *Microbiology* **147,** 519-26

Li, W. H., Wu, C. I. and Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2,** 150-74
Lin, J. and Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* **10,** 808-18
Liu, F. G., Miyamoto, M. M., Freire, N. P., Ong, P. Q., Tennant, M. R., Young, T. S. and Gugel, K. F. (2001). Molecular and morphological supertrees for eutherian (placental) mammals. *Science* **291,** 1786-9
Lobry, J. R. and Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res* **22,** 3174-80
Lopez, P., Philippe, H., Myllykallio, H. and Forterre, P. (1999). Identification of putative chromosomal origins of replication in Archaea. *Mol Microbiol* **32,** 883-6
Luria, S. E. and Human, M. (1952). A nonhereditary, host-induced variation of bacterial viruses. *J. Bact.* **64,** 557-569
Lusetti, S. L. and Cox, M. M. (2002). The bacterial reca protein and the recombinational DNA repair of stalled replication forks. *Annu Rev Biochem* **71,** 71-100
Macfadyen, L. P. (2000). Regulation of competence development in Haemophilus influenzae. *J Theor Biol* **207,** 349-59
MacFadyen, L. P., Chen, D., Vo, H. C., Liao, D., Sinotte, R. and Redfield, R. J. (2001). Competence development by Haemophilus influenzae is regulated by the availability of nucleic acid precursors. *Mol Microbiol* **40,** 700-7
MacNeill, S. A. (2001). Understanding the enzymology of archaeal DNA replication: progress in form and function. *Mol Microbiol* **40,** 520-9
Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. and Dowson, C. G. (2000). Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol* **182,** 1016-23
Marais, G. (2002). Les effets pervers du sexe sur l'évolution des génomes, Thése de l'université Lyon1 - Claude Bernard.
Margulis, L. (1970). *Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth.* Yale University Press.
Margulis, L. (1996). Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc Natl Acad Sci U S A* **93,** 1071-6
Martin, W. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* **21,** 99-104
Martin, W. and Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature* **392,** 37-41
Matte-Tailliez, O., Brochier, C., Forterre, P. and Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* **19,** 631-9
McGlynn, P. and Lloyd, R. G. (2002). Genome stability and the processing of damaged replication forks by RecG. *Trends Genet* **18,** 413-9
McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc Natl Acad Sci U S A* **95,** 10698-703
Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. *J Mol Biol* **222,** 851-6
Milkman, R. and Bridges, M. M. (1990). Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics* **126,** 505-17

Milkman, R. and Bridges, M. M. (1993). Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons. *Genetics* **133,** 455-68

Miller, S. L. and Lazcano, A. (1995). The origin of life--did it occur at high temperatures? *J Mol Evol* **41,** 689-92

Mira, A., Ochman, H. and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17,** 589-96

Moreira, D. and Lopez-Garcia, P. (1998). Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47,** 517-30

Moszer, I., Rocha, E. P. and Danchin, A. (1999). Codon usage and lateral gene transfer in Bacillus subtilis. *Curr Opin Microbiol* **2,** 524-8

Mulugu, S., Potnis, A., Shamsuzzaman, Taylor, J., Alexander, K. and Bastia, D. (2001). Mechanism of termination of DNA replication of Escherichia coli involves helicase-contrahelicase interaction. *Proc Natl Acad Sci U S A* **98,** 9569-74

Mushegian, A. R. and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93,** 10268-73

Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H. and Forterre, P. (2000). Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288,** 2212-5

Naya, H., Romero, H., Zavala, A., Alvarez, B. and Musto, H. (2002). Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. *J Mol Evol* **55,** 260-4

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., Fraser, C. M. and et al. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399,** 323-9

Nesbo, C. L., Boucher, Y. and Doolittle, W. F. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* **53,** 340-50

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Welti, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., Danson, M. J., Hough, D. W., Maddocks, D. G., Jablonski, P. E., Krebs, M. P., Angevine, C. M., Dale, H., Isenbarger, T. A., Peck, R. F., Pohlschroder, M., Spudich, J. L., Jung, K. W., Alam, M., Freitas, T., Hou, S., Daniels, C. J., Dennis, P. P., Omer, A. D., Ebhardt, H., Lowe, T. M., Liang, P., Riley, M., Hood, L. and DasSarma, S. (2000). Genome sequence of Halobacterium species NRC-1. *Proc Natl Acad Sci U S A* **97,** 12176-81

Nieselt-Struwe, K. and von Haeseler, A. (2001). Quartet-mapping, a generalization of the likelihood-mapping procedure. *Mol Biol Evol* **18,** 1204-19

O'Reilly, M. and Devine, K. M. (1997). Expression of AbrB, a transition state regulator from Bacillus subtilis, is growth phase dependent in a manner resembling that of Fis, the nucleoid binding protein from Escherichia coli. *J Bacteriol* **179,** 522-9

Ochman, H. (2001). Lateral and oblique gene transfer. *Curr Opin Genet Dev* **11,** 616-9

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299-304

Ochman, H. and Selander, R. K. (1984). Evidence for clonal population structure in Escherichia coli. *Proc Natl Acad Sci U S A* **81,** 198-201

Ochman, H., Soncini, F. C., Solomon, F. and Groisman, E. A. (1996). Identification of a pathogenicity island required for Salmonella survival in host cells. *Proc Natl Acad Sci U S A* **93,** 7800-4

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,** 907-30

Perals, K., Capiaux, H., Vincourt, J. B., Louarn, J. M., Sherratt, D. J. and Cornet, F. (2001). Interplay between recombination, cell division and chromosome structure during chromosome dimer resolution in Escherichia coli. *Mol Microbiol* **39,** 904-13

Perals, K., Cornet, F., Merlet, Y., Delon, I. and Louarn, J. M. (2000). Functional polarization of the Escherichia coli chromosome terminus: the dif site acts in chromosome dimer resolution only when located between long stretches of opposite polarity. *Mol Microbiol* **36,** 33-43

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A. and Blattner, F. R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409,** 529-33

Perriere, G., Bessieres, P. and Labedan, B. (2000a). EMGLib: the enhanced microbial genomes library (update 2000). *Nucleic Acids Res* **28,** 68-71

Perriere, G., Duret, L. and Gouy, M. (2000b). HOBACGEN: database system for comparative genomics in bacteria. *Genome Res* **10,** 379-85

Philippe, H. and Douzery, E. (1994). The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships. *J. Mam. Evol.* **2,** 133-152

Philippe, H. and Forterre, P. (1999). The rooting of the universal tree of life is not reliable. *J Mol Evol* **49,** 509-23

Radstrom, P., Fermer, C., Kristiansen, B. E., Jenkins, A., Skold, O. and Swedberg, G. (1992). Transformational exchanges in the dihydropteroate synthase gene of Neisseria meningitidis: a novel mechanism for acquisition of sulfonamide resistance. *J Bacteriol* **174,** 6386-93

Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* **1,** 53-8

Ragan, M. A. (2001). On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201,** 187-91

Rayssiguier, C., Thaler, D. S. and Radman, M. (1989). The barrier to recombination between Escherichia coli and Salmonella typhimurium is disrupted in mismatch-repair mutants. *Nature* **342,** 396-401

Redfield, R. J. (1993). Evolution of natural transformation: testing the DNA repair hypothesis in Bacillus subtilis and Haemophilus influenzae. *Genetics* **133,** 755-61

Redfield, R. J. (2001). Do bacteria have sex? *Nat Rev Genet* **2,** 634-9

Reeves, P. (1993). Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale. *Trends Genet* **9,** 17-22

Rivera, M. C., Jain, R., Moore, J. E. and Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95,** 6239-44

Rivera, M. C. and Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257,** 74-6

Roberts, R. J. and Macelis, D. (2000). REBASE - restriction enzymes and methylases. *Nucleic Acids Res* **28,** 306-7

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* **53,** 131-147
Rocha, E. P. and Blanchard, A. (2002). Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. *Nucleic Acids Res* **30,** 2031-42
Rocha, E. P. and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet* **18,** 291-4
Rocha, E. P., Danchin, A. and Viari, A. (1999a). Functional and evolutionary roles of long repeats in prokaryotes. *Res Microbiol* **150,** 725-33
Rocha, E. P., Danchin, A. and Viari, A. (1999b). Universal replication biases in bacteria. *Mol Microbiol* **32,** 11-6
Rocha, E. P., Danchin, A. and Viari, A. (2001). Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res* **11,** 946-58
Rocha, E. P., Matic, I. and Taddei, F. (2002). Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? *Nucleic Acids Res* **30,** 1886-94
Roger, A. J. and Brown, J. R. (1996). A chimeric origin for eukaryotes re-examined. *Trends Biochem Sci* **21,** 370-2
Rossello-Mora, R. and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25,** 39-67
Sawitzke, J. and Austin, S. (2001). An analysis of the factory model for chromosome replication and segregation in bacteria. *Mol Microbiol* **40,** 786-94
Schneider, R., Lurz, R., Luder, G., Tolksdorf, C., Travers, A. and Muskhelishvili, G. (2001). An architectural role of the Escherichia coli chromatin protein FIS in organising DNA. *Nucleic Acids Res* **29,** 5107-14
Schwartz, R. M. and Dayhoff, M. O. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199,** 395-403
Schwartz, R. M. and Dayhoff, M. O. (1981). Chloroplast origins: inferences from protein and nucleic acid sequences. *Ann N Y Acad Sci* **361,** 260-72
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,** 1281-95
Sharp, P. M., Shields, D. C., Wolfe, K. H. and Li, W. H. (1989). Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246,** 808-10
She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P. M., Heikamp-de Jong, I., Jeffries, A. C., Kozera, C. J., Medina, N., Peng, X., Thi-Ngoc, H. P., Redder, P., Schenk, M. E., Theriault, C., Tolstrup, N., Charlebois, R. L., Doolittle, W. F., Duguet, M., Gaasterland, T., Garrett, R. A., Ragan, M. A., Sensen, C. W. and Van der Oost, J. (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proc Natl Acad Sci U S A* **98,** 7835-40
Shields, D. C. and Sharp, P. M. (1989). Evidence that mutation patterns vary among Drosophila transposable elements. *J Mol Biol* **207,** 843-6
Simpson, W. J., Musser, J. M. and Cleary, P. P. (1992). Evidence consistent with horizontal transfer of the gene (emm12) encoding serotype M12 protein between group A and group G pathogenic streptococci. *Infect Immun* **60,** 1890-3
Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147,** 195-7
Snel, B., Bork, P. and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet* **21,** 108-10

Sohngen, N. L. (1906). Ueber Bakterien, welche Methan als Kohlenstoffnahrung und energiequelle gebrauchen. *Zentralbl. Bakteriol. Parasitik. Abt. I.* **15,** 513-517
Spratt, B. G., Bowler, L. D., Zhang, Q. Y., Zhou, J. and Smith, J. M. (1992). Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species. *J Mol Evol* **34,** 115-25
Srivastava, A. K. and Schlessinger, D. (1990). Preparation of extracts and assay of ribosomal RNA maturation in Escherichia coli. *Methods Enzymol* **181,** 355-66
Strauss, B. S. (1991). The 'A rule' of mutagen specificity: a consequence of DNA polymerase bypass of non-instructional lesions? *Bioessays* **13,** 79-84
Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13,** 964-9
Strimmer, K. and von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A* **94,** 6815-9
Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* **48,** 582-592
Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85,** 2653-7
Sueoka, N. (1992). Directional mutation pressure, selective constraints, and genetic equilibria. *J Mol Evol* **34,** 95-114
Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40,** 318-25
Sueoka, N. (1999). Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *J Mol Evol* **49,** 49-62
Sumi, M., Sato, M. H., Denda, K., Date, T. and Yoshida, M. (1992). A DNA fragment homologous to F1-ATPase beta subunit was amplified from genomic DNA of Methanosarcina barkeri. Indication of an archaebacterial F-type ATPase. *FEBS Lett* **314,** 207-10
Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* **28,** 237-61
Syvanen, M., Hartman, H. and Stevens, P. F. (1989). Classical plant taxonomic ambiguities extend to the molecular level. *J Mol Evol* **28,** 536-44
Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28,** 33-6
Teichmann, S. A. and Mitchison, G. (1999). Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol* **49,** 98-107
Tekaia, F., Lazcano, A. and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res* **9,** 550-7
Tenaillon, O., Taddei, F., Radmian, M. and Matic, I. (2001). Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res Microbiol* **152,** 11-6
Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F. and Godelle, B. (1999). Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics* **152,** 485-93
Thanaraj, T. A. and Argos, P. (1996a). Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci* **5,** 1973-83
Thanaraj, T. A. and Argos, P. (1996b). Ribosome-mediated translational pause and protein domain organization. *Protein Sci* **5,** 1594-612
Thioulouse, J., Chessel, D., Dolédec, S. and Olivier, J. M. (1997). ADE-4: a multivariate analysis and graphical display software. *Stat. Comput* **7,** 75-83

Thomarat, F. (2002). Analyse phylogénétique du génome complet de la microsporidie Encephalitozoon cuniculi, Thése de l'université Lyon1-Claude Bernard.
Travers, A., Schneider, R. and Muskhelishvili, G. (2001). DNA supercoiling and transcription in Escherichia coli: The FIS connection. *Biochimie* **83,** 213-7
Tsai, L. and Sun, Z. (2001). Dynamic flexibility in the Escherichia coli genome. *FEBS Lett* **507,** 225-30
Tsutsumi, S., Denda, K., Yokoyama, K., Oshima, T., Date, T. and Yoshida, M. (1991). Molecular cloning of genes encoding major two subunits of a eubacterial V-type ATPase from Thermus thermophilus. *Biochim Biophys Acta* **1098,** 13-20
Ussery, D., Larsen, T. S., Wilkes, K. T., Friis, C., Worning, P., Krogh, A. and Brunak, S. (2001). Genome organisation and chromatin structure in Escherichia coli. *Biochimie* **83,** 201-12
Wainwright, M. (1997). Extreme pleomorphism and the bacterial life cycle: a forgotten controversey. *Perspectives in Biology and Medicine* **40,** 407-414
Wake, R. G. (1997). Replication fork arrest and termination of chromosome replication in Bacillus subtilis. *FEMS Microbiol Lett* **153,** 247-54
Wang, B. (2001). Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol* **53,** 244-50
Whatmore, A. M. and Kehoe, M. A. (1994). Horizontal gene transfer in the evolution of group A streptococcal emm-like genes: gene mosaics and variation in Vir regulons. *Mol Microbiol* **11,** 363-74
White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., Fraser, C. M. and et al. (1999). Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. *Science* **286,** 1571-7
Whittam, T. S., Ochman, H. and Selander, R. K. (1983). Multilocus genetic structure in natural populations of Escherichia coli. *Proc Natl Acad Sci U S A* **80,** 1751-5
Woese, C. (1987). Bacterial evolution. *Microbiol. Rev.* **51,** 221-271
Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74,** 5088-90
Woese, C. R., Stackebrandt, E. and Ludwig, W. (1984). What are mycoplasmas: the relationship of tempo and mode in bacterial evolution. *J Mol Evol* **21,** 305-16
Woese, C. R., Stackebrandt, E., Macke, T. J. and Fox, G. E. (1985). A phylogenetic definition of the major eubacterial taxa. *Syst Appl Microbiol* **6,** 143-51
Woldringh, C. L. (2002). The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol Microbiol* **45,** 17-29
Woldringh, C. L., Jensen, P. R. and Westerhoff, H. V. (1995). Structure and partitioning of bacterial DNA: determined by a balance of compaction and expansion forces? *FEMS Microbiol Lett* **131,** 235-42
Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. and Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1,** 8
Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Applic. Biosci.* **13,** 555-556
Yoder, A. D., Irwin, J. A. and Payseur, B. A. (2001). Failure of the ILD to determine data combinability for slow loris phylogeny. *Syst Biol* **50,** 408-24

Zhaxybayeva, O. and Gogarten, J. P. (2002). Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* **3,** 4

Zillig, W. (1987). Eukaryotic traits in Archaebacteria. Could the eukaryotic cytoplasm have arisen from archaebacterial origin? *Ann N Y Acad Sci* **503,** 78-82

Zillig, W., Schnabel, R. and Stetter, K. O. (1985). Archaebacteria and the origin of the eukaryotic cytoplasm. *Curr Top Microbiol Immunol* **114,** 1-18

Zinder, N. and Lederberg, J. (1952). Genetic exchange in Salmonella. *J. Bact.* **64,** 679-699

Zivanovic, Y., Lopez, P., Philippe, H. and Forterre, P. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* **30,** 1902-10

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theoret Biol* **8,** 357-66

Zuckerman, H. and Lederberg, J. (1986). Forty years of genetic recombination in bacteria. Postmature scientific discovery? *Nature* **324,** 629-631