(Ré)annotation de génomes procaryotes

Exploration de groupes de gènes chez les bactéries

Stéphanie BOCS

Atelier de Génomique Comparative

19 mai 2004

Plan

I. Annotation de génomes procaryotes

II. Développements méthodologiques

III. Exploration de groupes de gènes

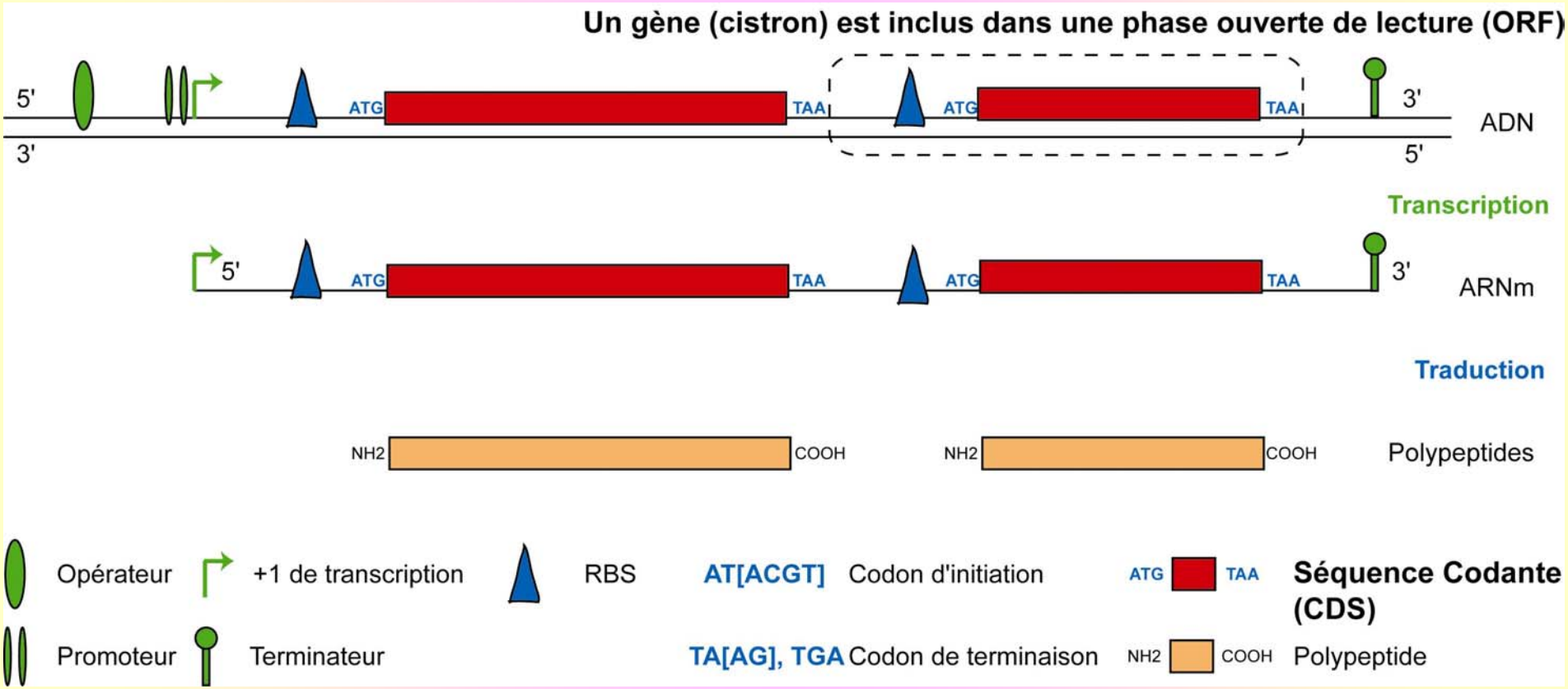
IV. Conclusion

Avancée du séquençage et de l'annotation des génomes en Avril 2004

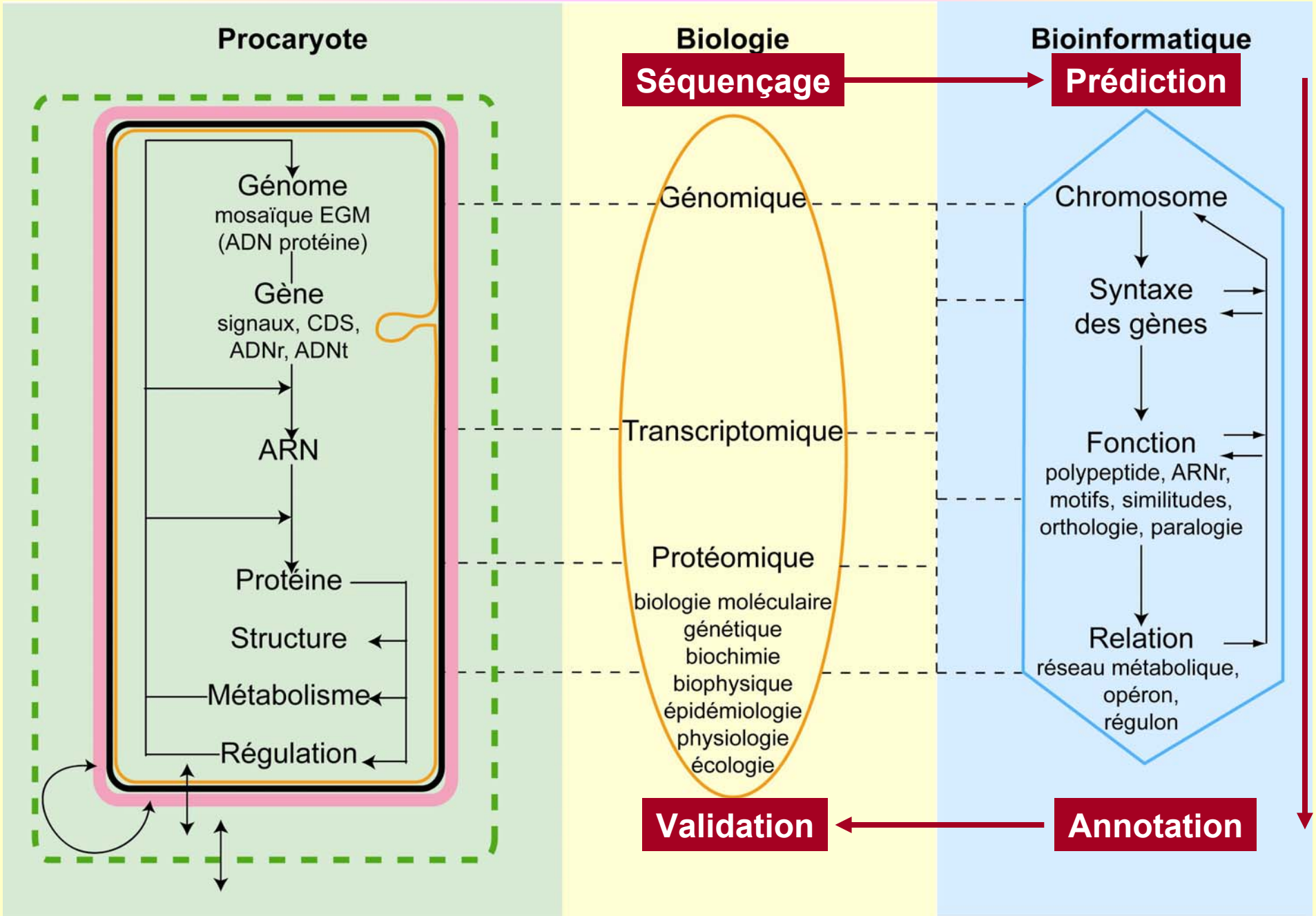
	Procaryotes			
Projets	Archées	Bactéries	Eucaryotes	Total
Publiés	18	142	26	186
En cours	26	461	415	902

- Organisation génomique commune entre Archées et Bactéries
- Les procaryotes ont un génome compact et révèlent une grande biodiversité
- Étudier la biodiversité nécessite d'annoter les séquences biologiques

Opéron codant des polypeptides



Niveaux d'annotation et d'exploration

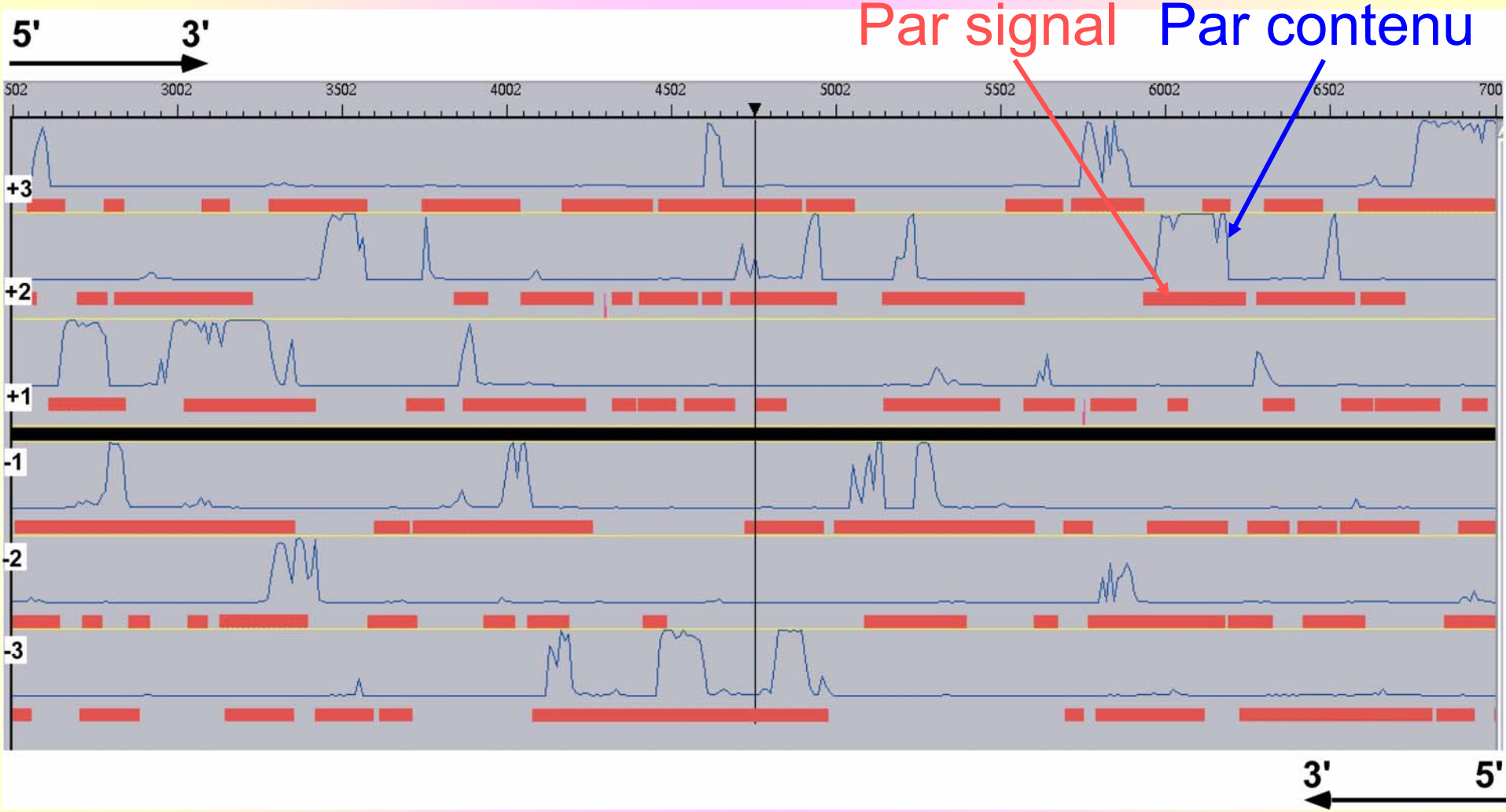


Méthodes de prédiction de CDS

- Extrinsèques

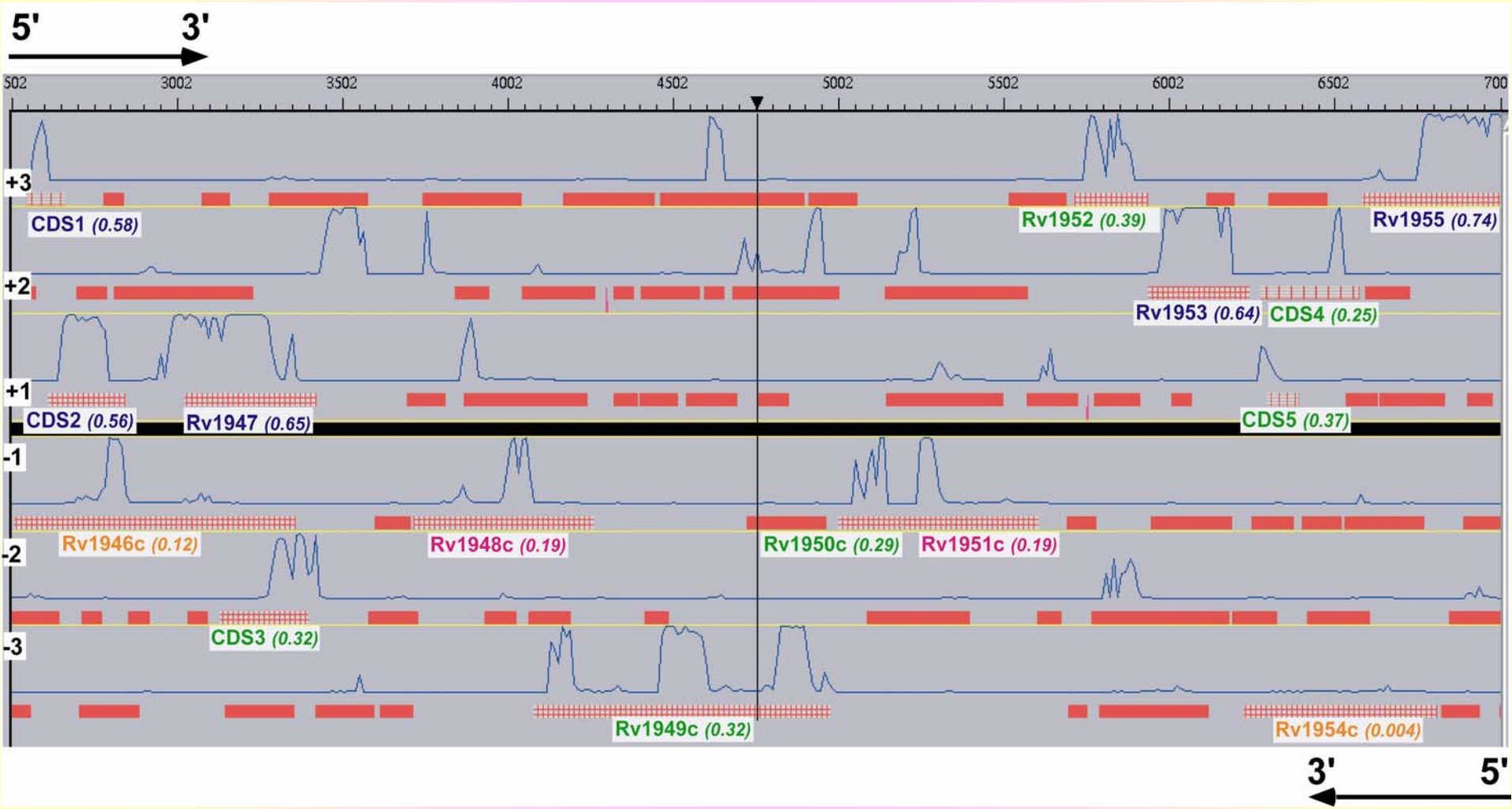
  recherche de similitudes dans les banques par alignement de paires de séquences
- Intrinsèques
  - Par signal
  - Par contenu
    - – Sans modèle d'ADN codant
    - – Avec un modèle d'ADN codant

Prédiction automatique de CDS sur un fragment du chromosome de *Mycobacterium tuberculosis*



Imagene [Médigue C. *et al.* 1999 Bioinformatics]

Annotation manuelle de CDS sur le fragment de *M. tuberculosis*



Anomalies de CDS

Observations de fragments:

- Décalage du cadre de lecture des CDS (***frameshift***)
- Absence ou présence de codon(s) stop en phase
- Délétion ou duplication partielle

Origine:

- Erreur de séquençage
- Anomalie authentique

  Mécanismes:

  – Mutation ponctuelle d'une ou plusieurs pb

  – Recombinaison homologue (conjugaison, transformation)

  – Recombinaison illégitime (transfert horizontal par transduction, transposition de séquences d'insertion (**IS**))

  Fonction:

  – Programmée

  – **Pseudogène**

Hétérogénéité des annotations ou biodiversité ?

Division	Nom commun	Code espèce	Génome		Nombre		
			Taille	G+C	CDS	Ψ	IS
Protéobactérie γ	*Escherichia coli*	ECOLI	4,64	50,8	4311	33	42
		ECO57	5,53	50,4	5324	2	49
	Yersinia pestis	YERPE	4,65	47,6	4008	149	140
	Y. pseudotuberculosis	YERPS	4,74		3977	62	20
	Photorhabdus luminescens subsp. laumondii	PHOLL	5,69	42,8	4905	222	195
Protéobactérie β	*N. meningitidis*	NEIMA	2,18	51,8	2121	56	62
		NEIMB	2,27	51,5	2158	79	51
Firmicute	*M. tuberculosis*	MYCTC	4,40	65,6	4187	35	33
		MYCTU	4,41		3927	55	58
	B. halodurans	BACHD	4,20	43,7	4066	1	120
	B. subtilis subsp. subtilis	BACSU	4,21	43,5	4112	1	NR

Ψ: pseudogène, **NR**: non rapporté.

Nécessité d'une stratégie de prédiction de CDS pour :

- annoter la séquence d'un projet en cours
- évaluer la « justesse » des annotations d'un projet public

Problématique de la thèse

- **Développements**
  - Stratégie de prédiction de CDS d'un génome
  - Système de gestion de base de données relationnelle (SGBDR)
  - Processus de réannotation de génomes publics
- **(Ré)annotations**
  - Génomes publics
  - Nouveaux génomes
- **Exploration** comparative d'annotations tels que les îlots génomiques[1]

[1] groupes de gènes probablement acquis par transfert horizontal

Pourquoi une nouvelle stratégie de prédiction de CDS ?

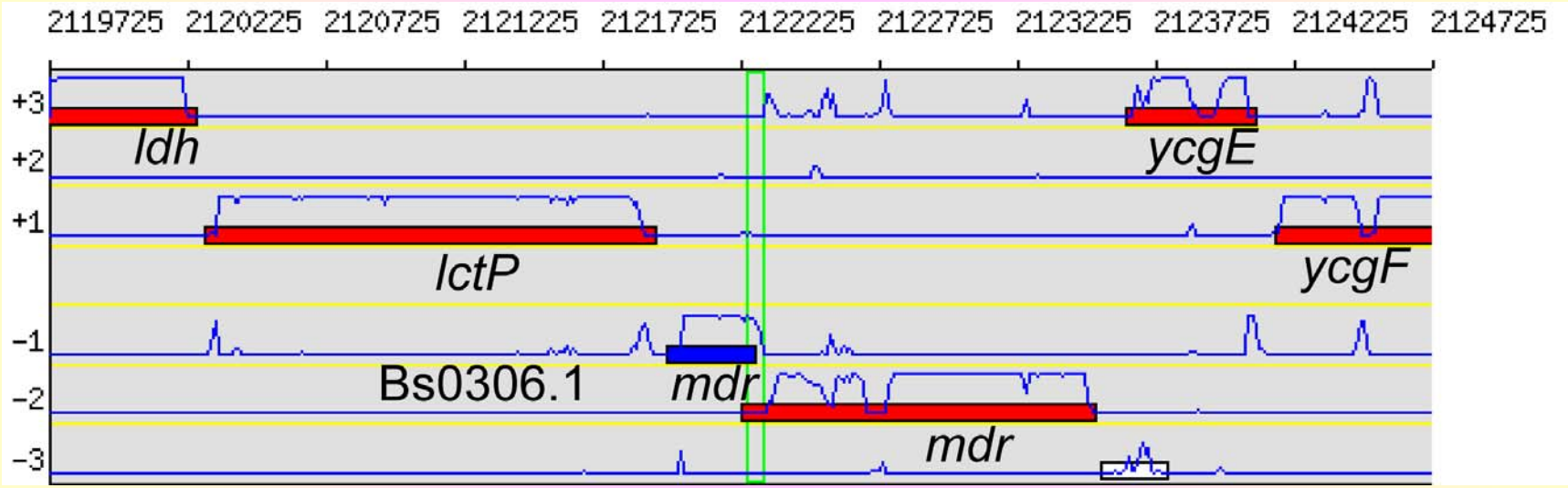
En 1999, les programmes **GeneMark**[1] et **Glimmer**[2] ne géraient pas correctement les:

- petites CDS (60 < L < 300 pb)
- petits recouvrements naturels entre CDS (< 30 pb)
- grands recouvrements artificiels (Ex 1,2)
- CDS de composition atypique (Ex 3)
- CDS fantômes

⇒ Problèmes accentués chez les génomes G+C riches

1 [Borodovsky M. & McIninch J. 1993 Computers Chem; Lukashin A. & Borodovsky M. 1998 Nucleic Acids Res]
2 [Delcher A. L. *et al.* 1999 Nucleic Acids Res]

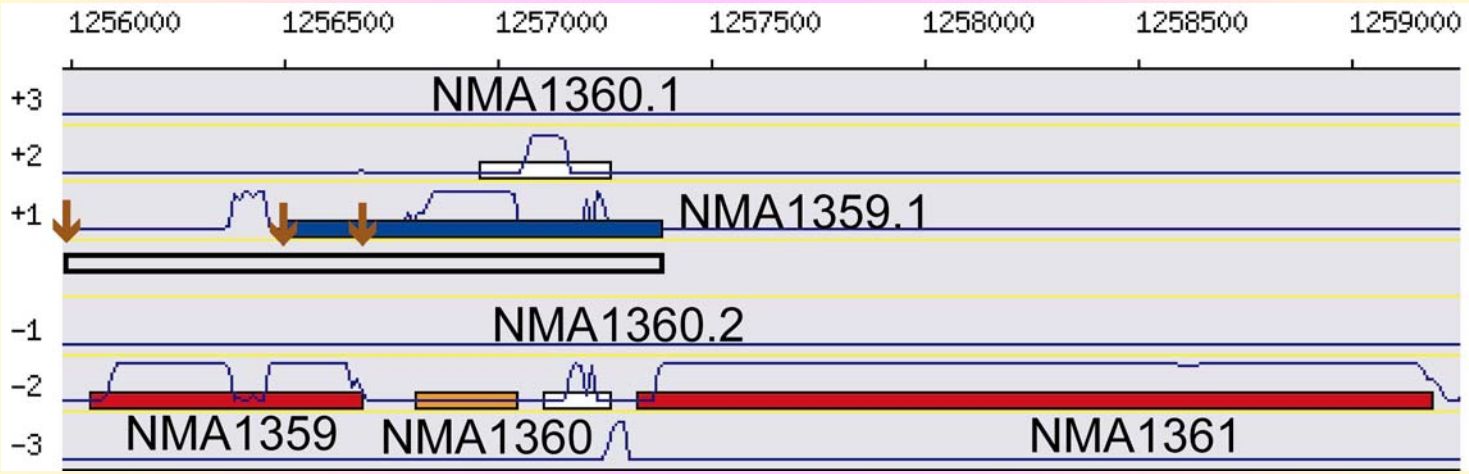
Ex 1 : *Frameshift* sur un fragment du chromosome de *B. subtilis*



➡ Chevauchement des deux fragments codant un transporteur *multidrug-efflux* de 51 pb

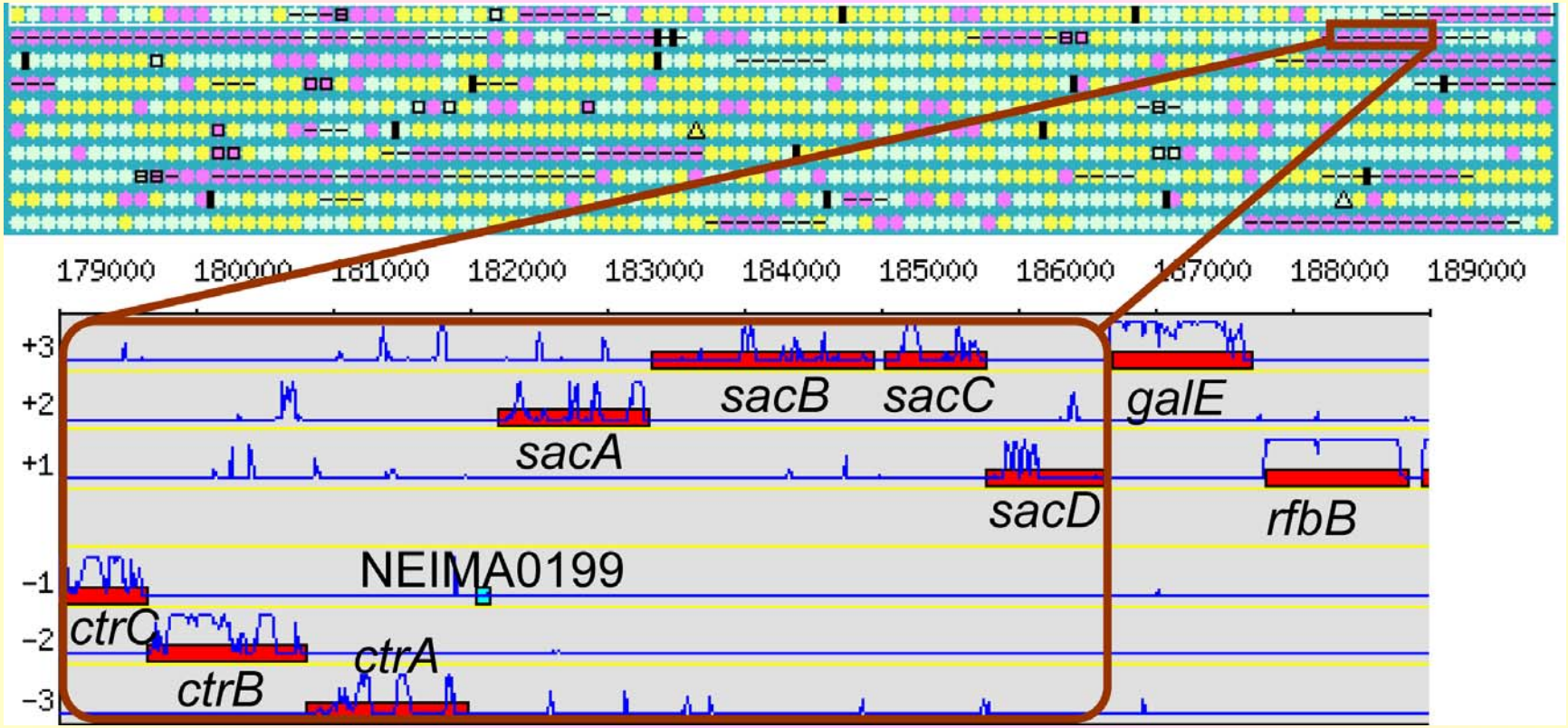
MaGe [Vallenet *et al.* (en préparation)]

Ex 2 : Codon d'initiation trop en 5' sur un fragment du chromosome de *N. meningitidis*



NMA1359, NMA1359.1 et NMA1360.1 codent des polypeptides de biosynthèse de la capsule

Ex 3 : Région atypique du chromosome de *N. meningitidis*



Hétérogénéité de composition des CDS:

la région atypique (A+T riches) impliquée dans la biosynthèse de la capsule est plus difficile à prédire (transfert horizontal ?) [Hsiao W. *et al.* 2003 Bioinformatics]

Stratégie experte semi-automatique de prédiction de CDS

- Stratégie d'apprentissage des séquences codantes et non-codantes d'un génome: **AMIMat** (K Matrices pour AMIGene)
- Stratégie de reconnaissance des CDS d'un génome: **AMIGene** (Annotation des gènes MIcrobiens)

S. Bocs

  - **Optimisations** empiriques et automatiques des **paramètres** d'AMIGene
  - Comparaison de la **précision** d'AMIGene avec d'autres programmes
  - **Application Web** AMIGene

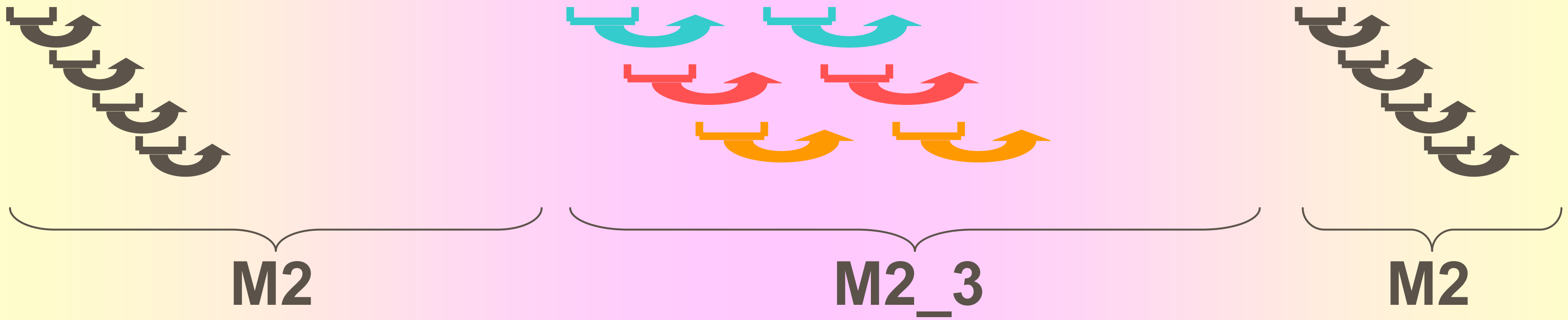
→ D. Vallenet

http://www.genoscope.cns.fr/agc/tools/amigene/

Principe des chaînes de Markov

$X_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots X_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots X_n$

$AGGAGGATTACCCCA_1T_2G_3G_1C_2T_3T_1A_2T_3T_1A_2A_3ATCTATTC$



$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{+1}) = \mu_2(A_1T_2)\pi_3(A_1T_2,G_3)\pi_1(T_2G_3,G_1)\pi_2(G_3G_1,C_2)\ldots$$

$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{+2}) = \ldots$$

$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{+3}) = \ldots$$

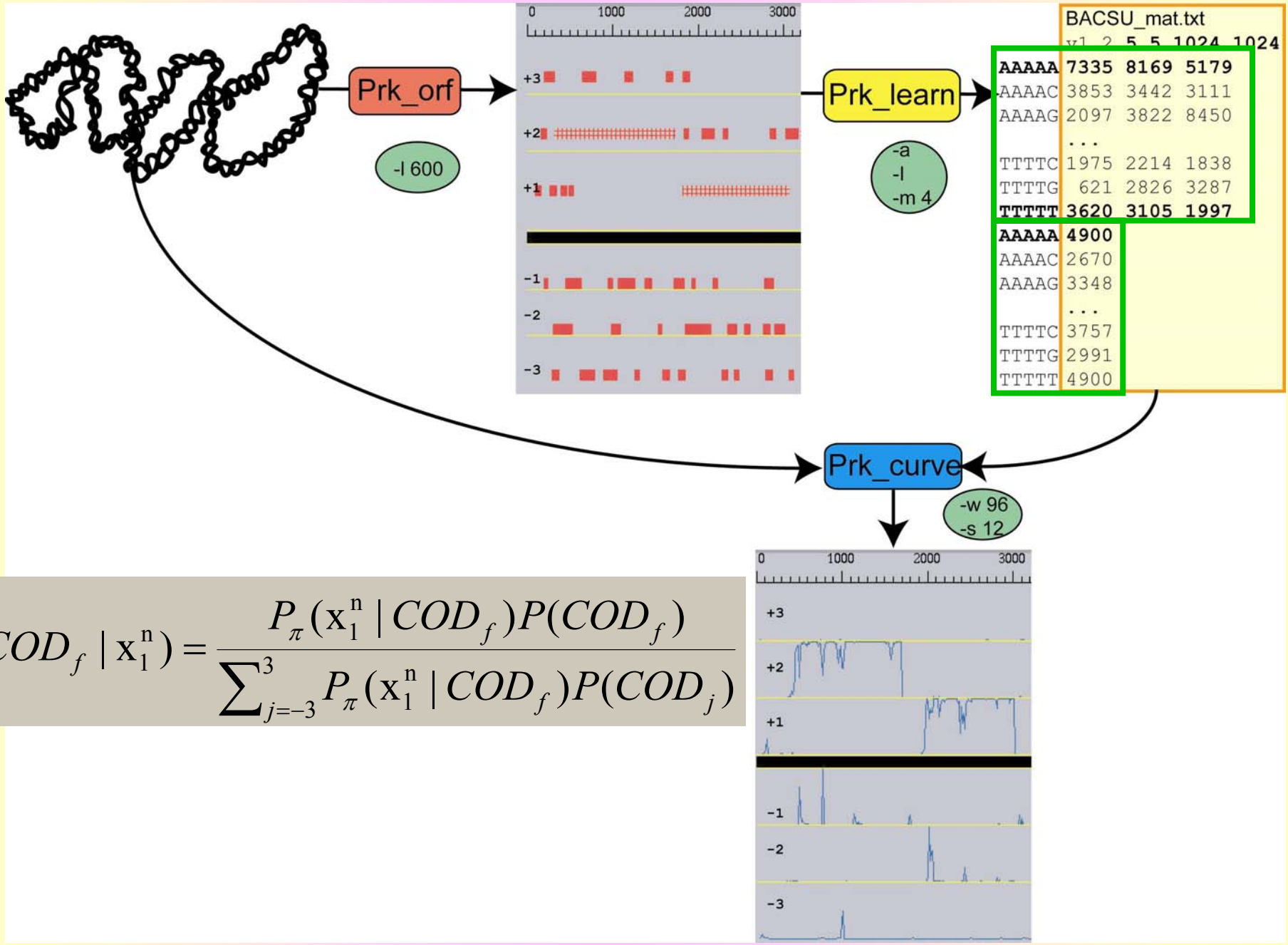
$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{0}) = \ldots$$

$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{-1}) = \ldots$$

$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{-2}) = \ldots$$

$$P_\pi(\mathrm{x}_1^n \mid \mathrm{COD}_{-3}) = \ldots$$

Modules Prokov



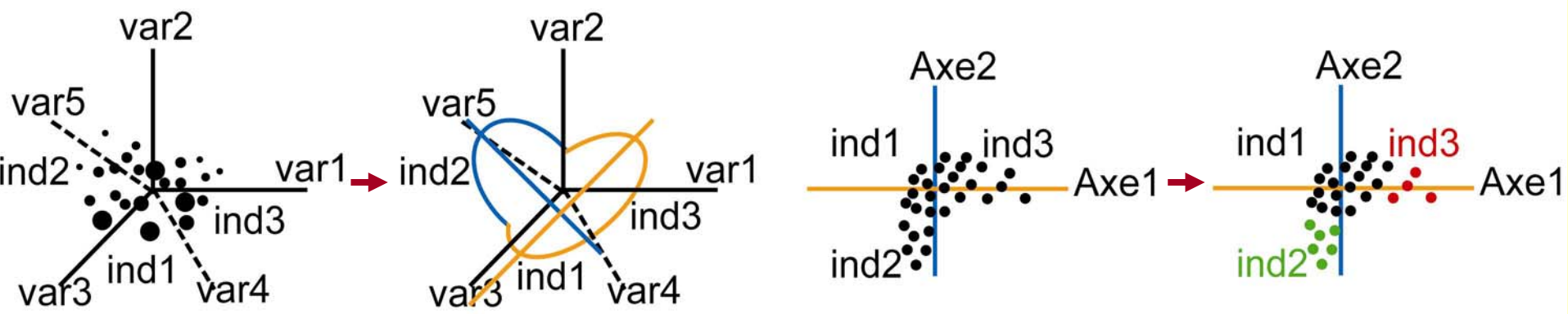
$$P_{\pi}(COD_f \mid \mathrm{x}_1^{\mathrm{n}}) = \frac{P_{\pi}(\mathrm{x}_1^{\mathrm{n}} \mid COD_f)P(COD_f)}{\sum_{j=-3}^{3} P_{\pi}(\mathrm{x}_1^{\mathrm{n}} \mid COD_f)P(COD_j)}$$

Principe des analyses multivariées

- **Fréquence Relative des Codons Synonymes** (RSCU)
  - biais dans l'usage des codons synonymes des CDS

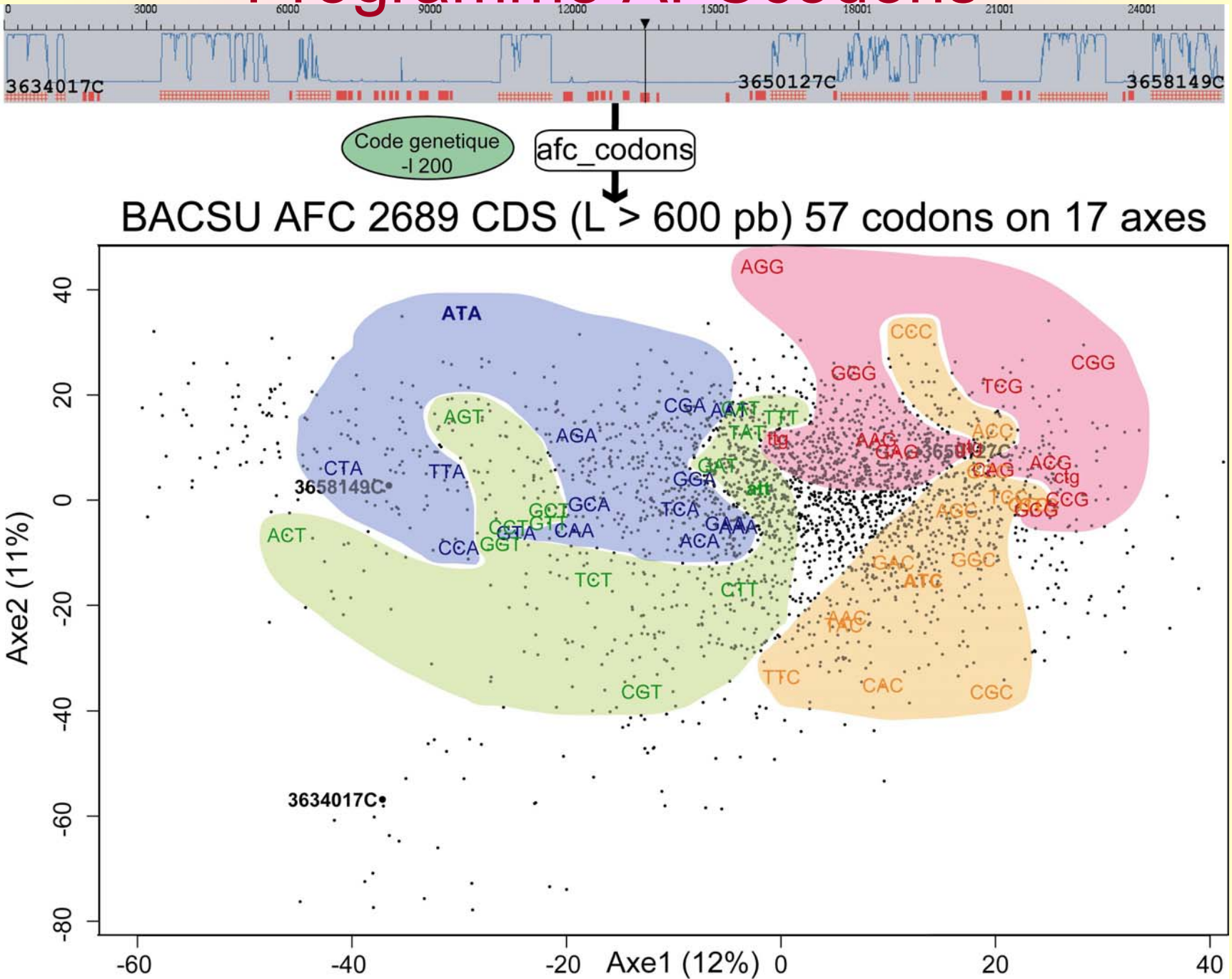
$$RSCU(abc) = \frac{N_{abc}}{\sum N_{a'b'c'}} Nsyn_{abc}$$

- **Analyse Factorielle des Correspondances** (AFC)
  - Tendances principales dans l'usage des codons synonymes des CDS
- **Centres mobiles** (K-means, Nuées dynamiques)
  - Classes de génes

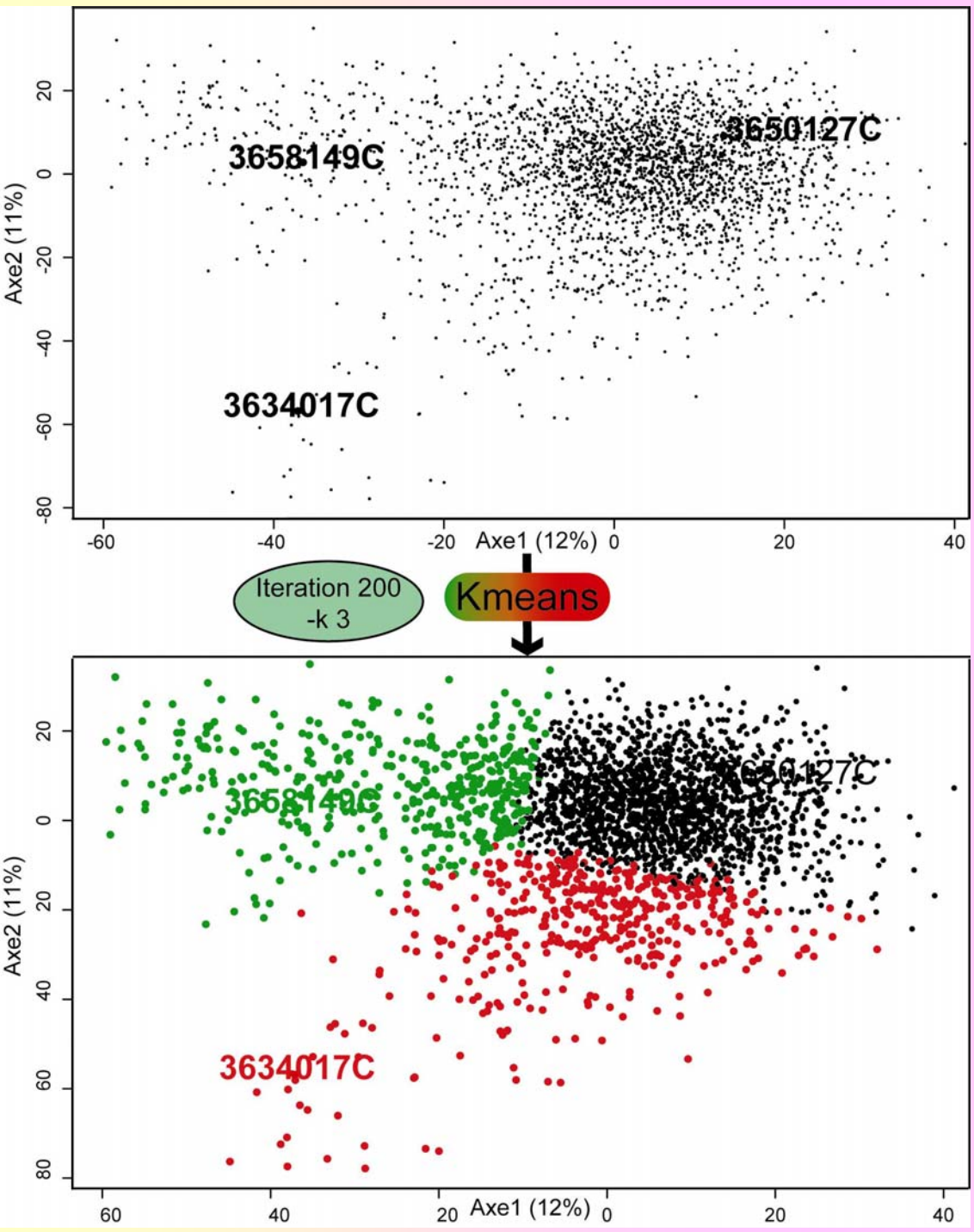


RSCU AFC K-means

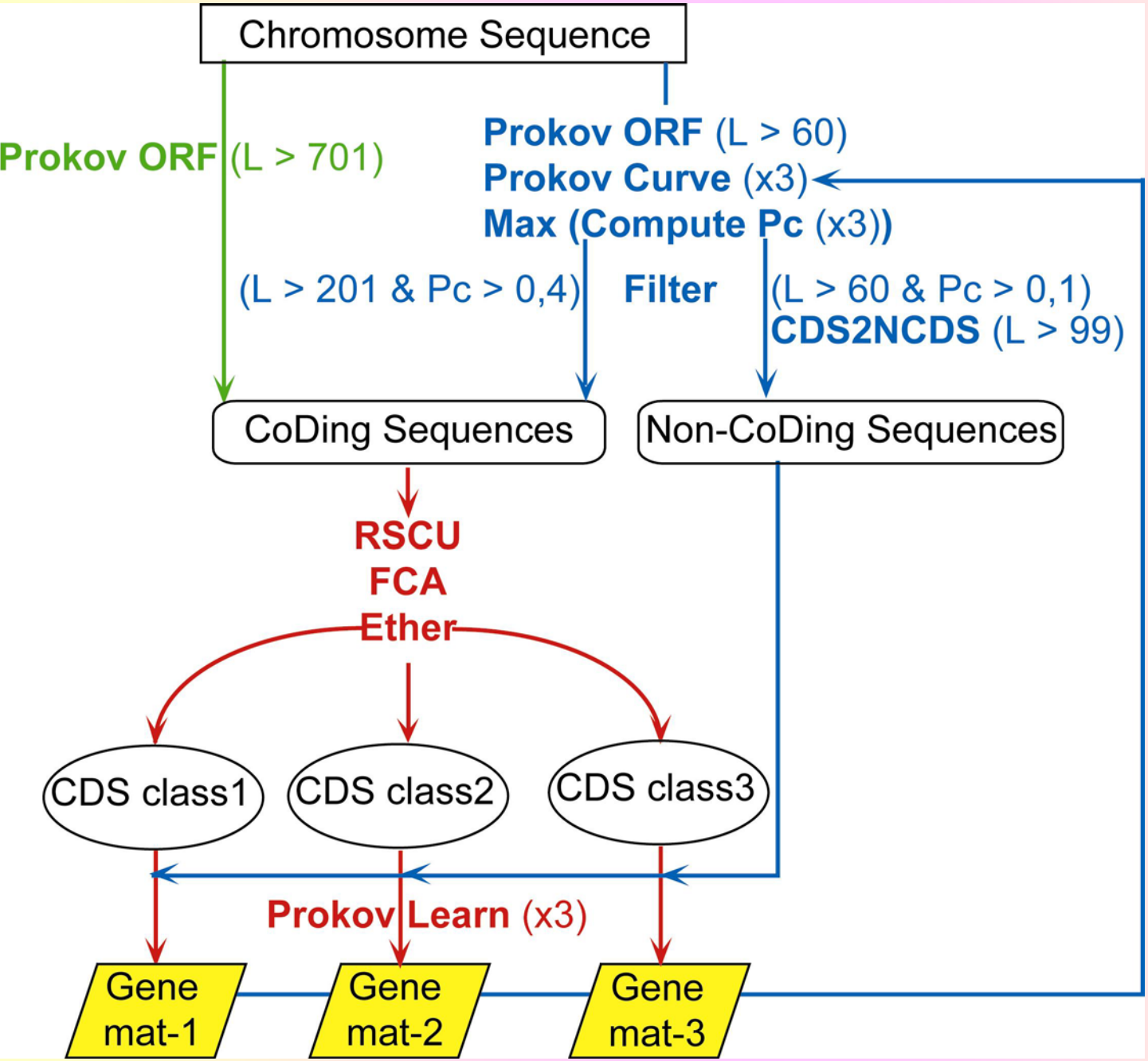
Programme AFCcodons



Programme K-means



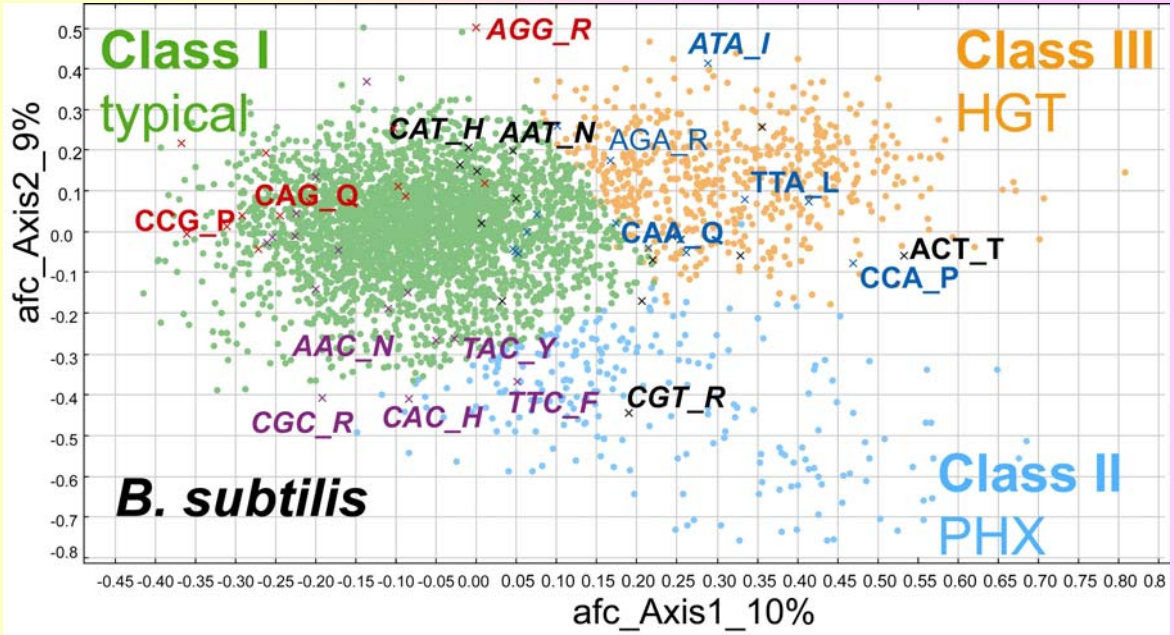
AMIMat: construction de matrices de transition



L: longueur en pb,
Pc: probabilité moyenne de codage d'une CDS.

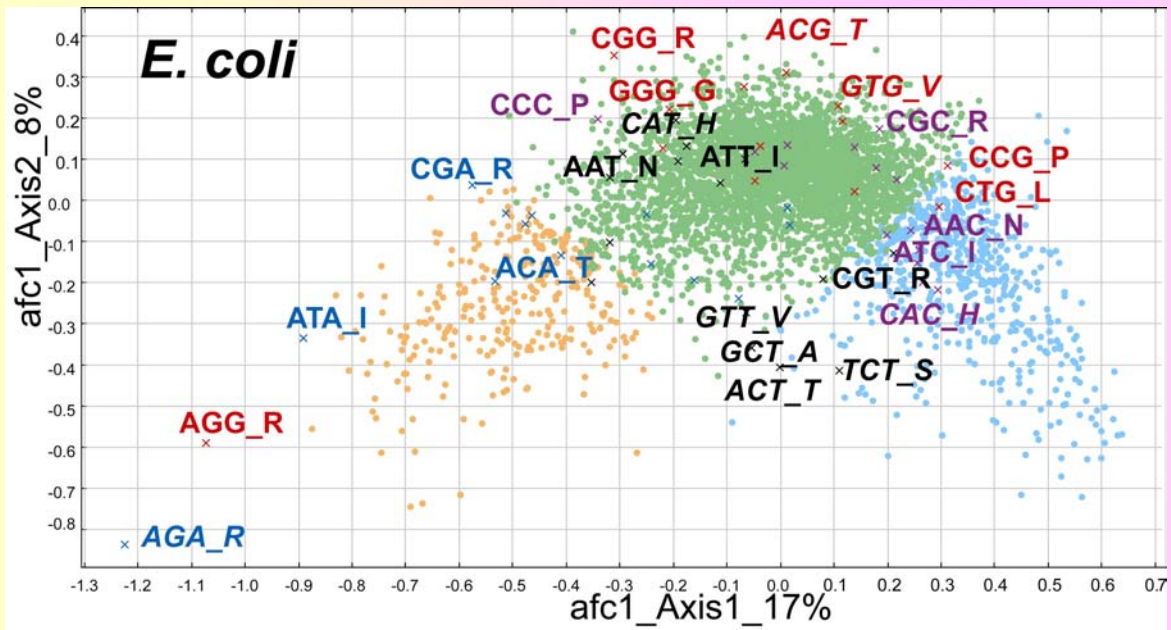
Apprentissage semi-automatique expert de CDS de composition hétérogène et de séquences non codantes

3 classes de gènes chez *B. subtilis* & *E. coli*



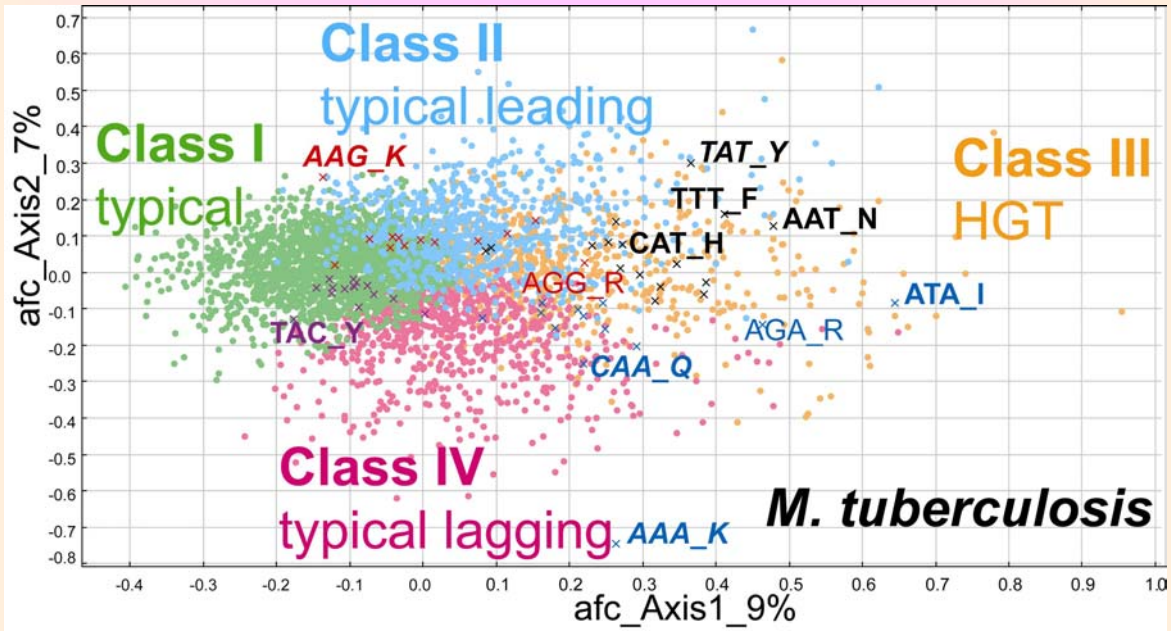
Genostar [Durand P. *et al.* 2003 Curr Opin Drug Discov]

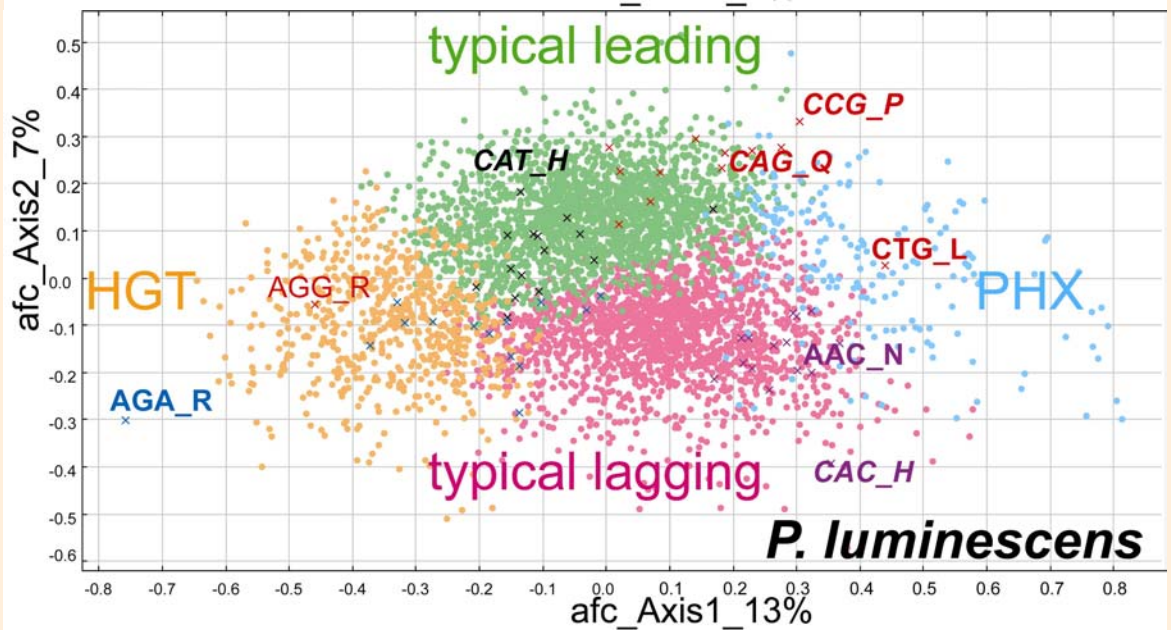
Moszer I. *et al.* (1999 Curr Opin Microbiol)



Médigue C. *et al.* (1991 J Mol Bio)

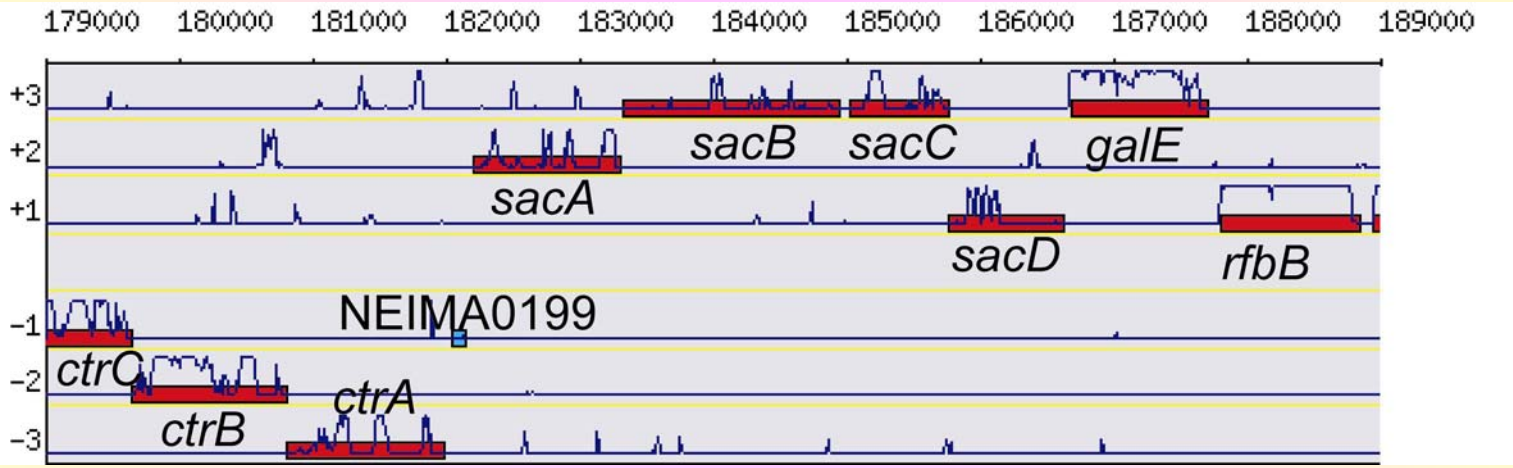
4 classes de gènes chez *M. tuberculosis* & *P. luminescens* ?



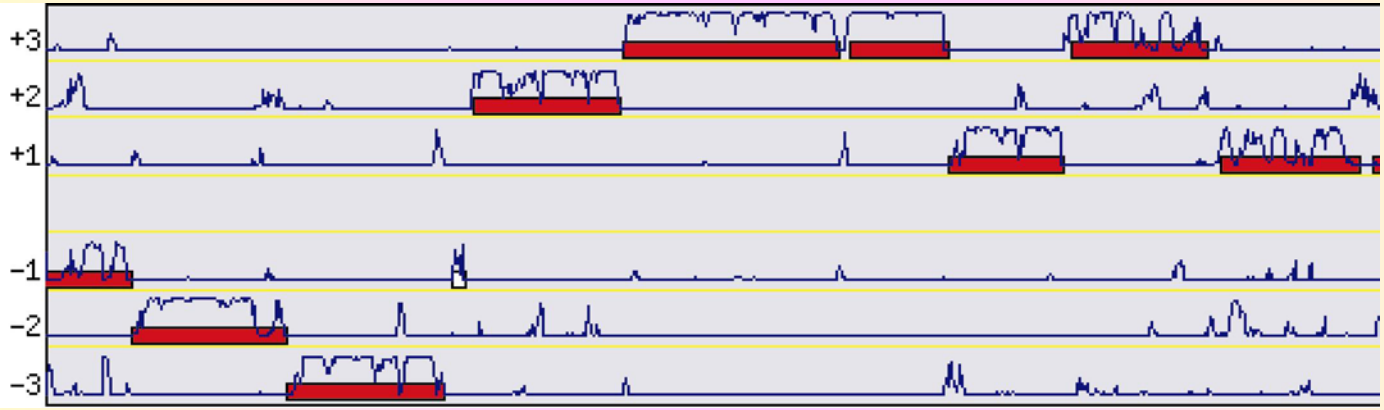


K matrices pour prendre en compte l'hétérogénéité des CDS

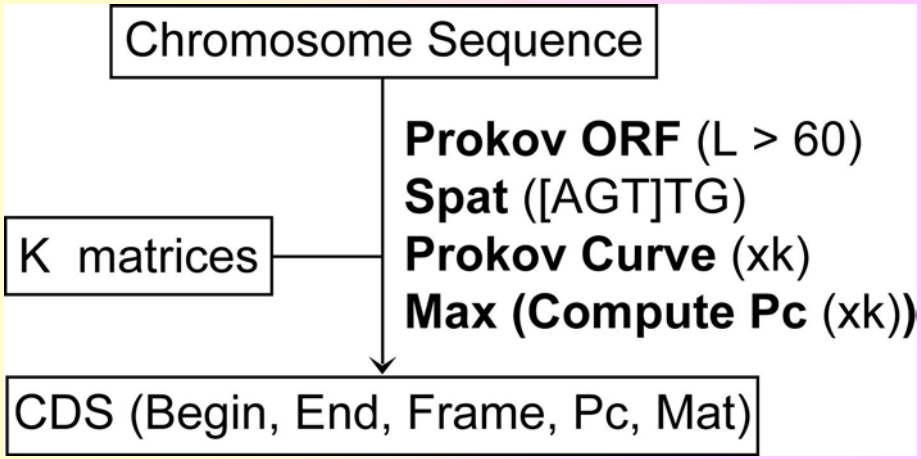
Matrice des CDS de classe I

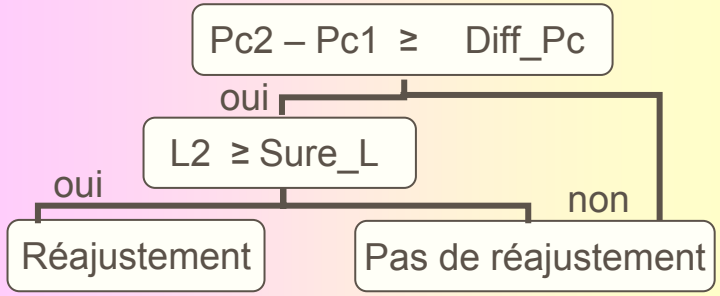


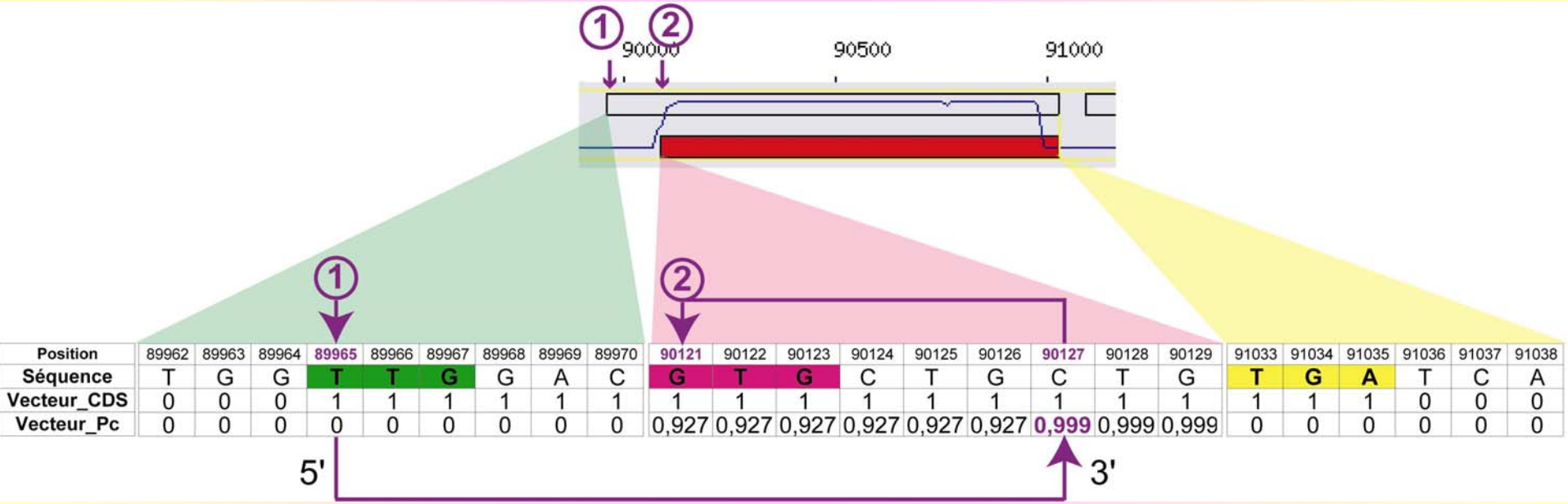
Matrice des CDS de classe III



AMIGene – phase I – reconnaissance

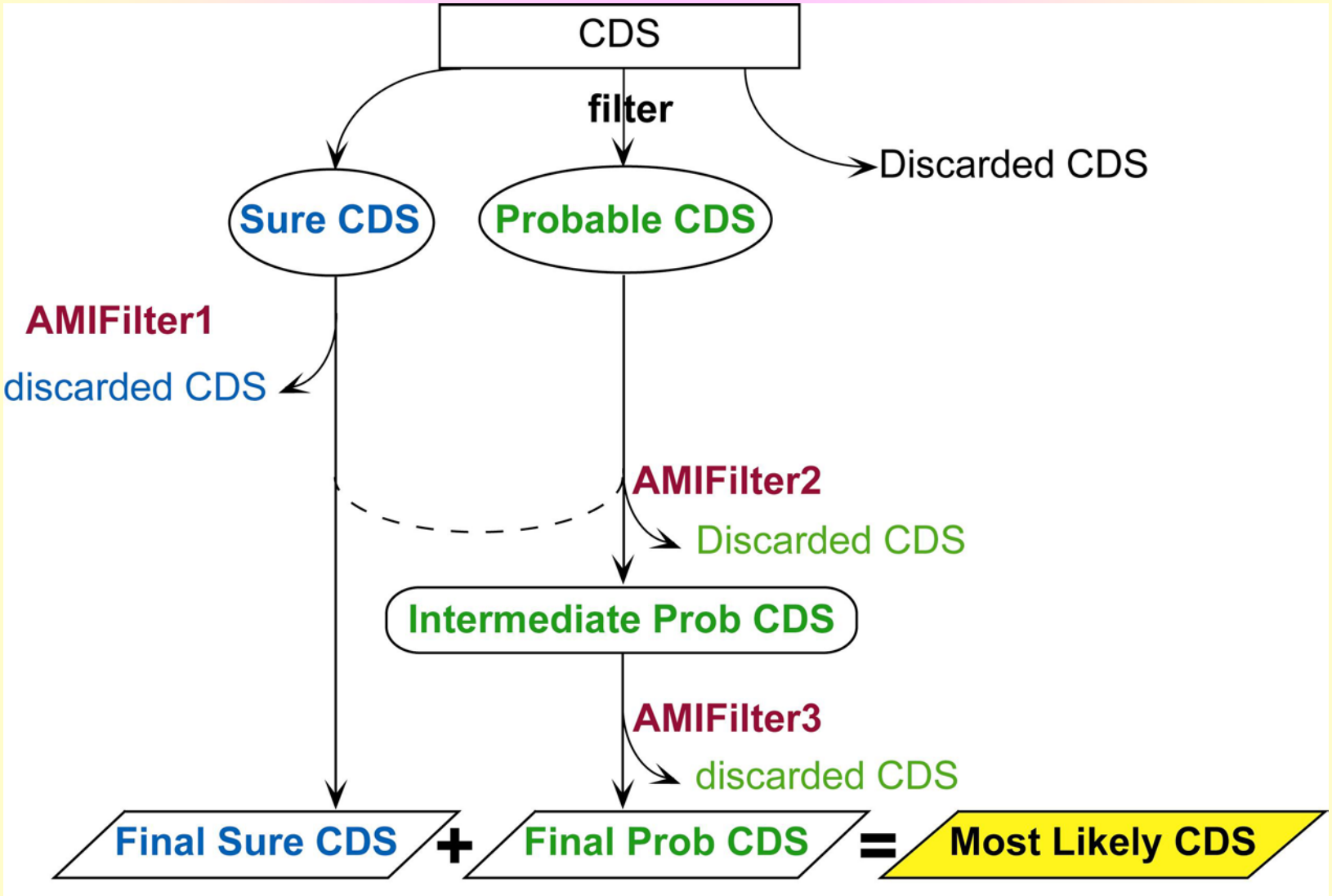




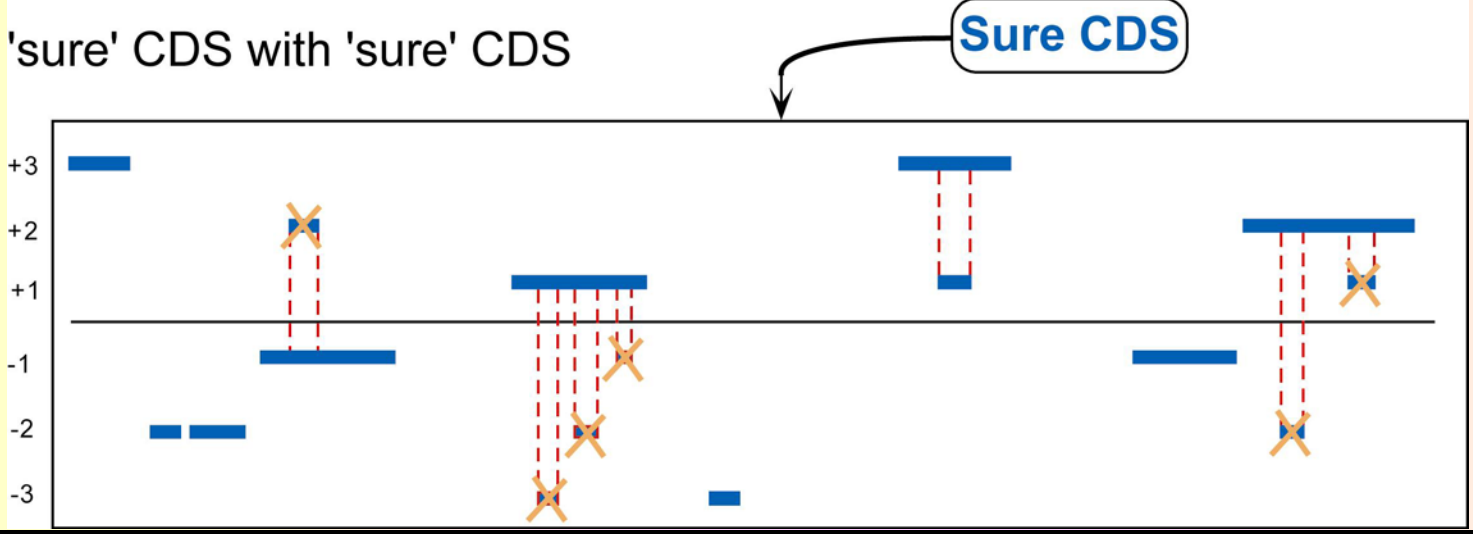


Position	89962	89963	89964	89965	89966	89967	89968	89969	89970	90121	90122	90123	90124	90125	90126	90127	90128	90129	91033	91034	91035	91036	91037	91038
Séquence	T	G	G	T	T	G	G	A	C	G	T	G	C	T	G	C	T	G	T	G	A	T	C	A
Vecteur_CDS	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
Vecteur_Pc	0	0	0	0	0	0	0	0	0	0,927	0,927	0,927	0,927	0,927	0,927	0,999	0,999	0,999	0	0	0	0	0	0

AMIGene – phase II – post-traitements



AMIFilter1

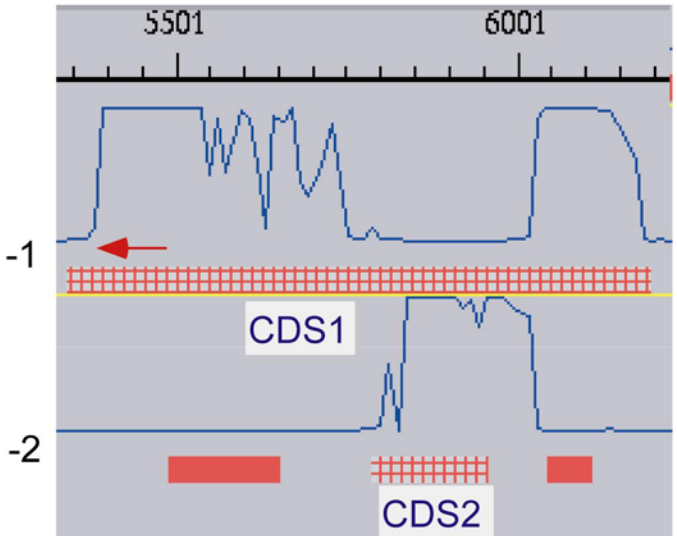


Bactérie	Nom	Début (brin)	L	Pc	% Inclusion	Résultat
E. coli	CDS1	5337 (r)	855	0.51		
	CDS2	5780 (r)	174	0.77	2O	conservée

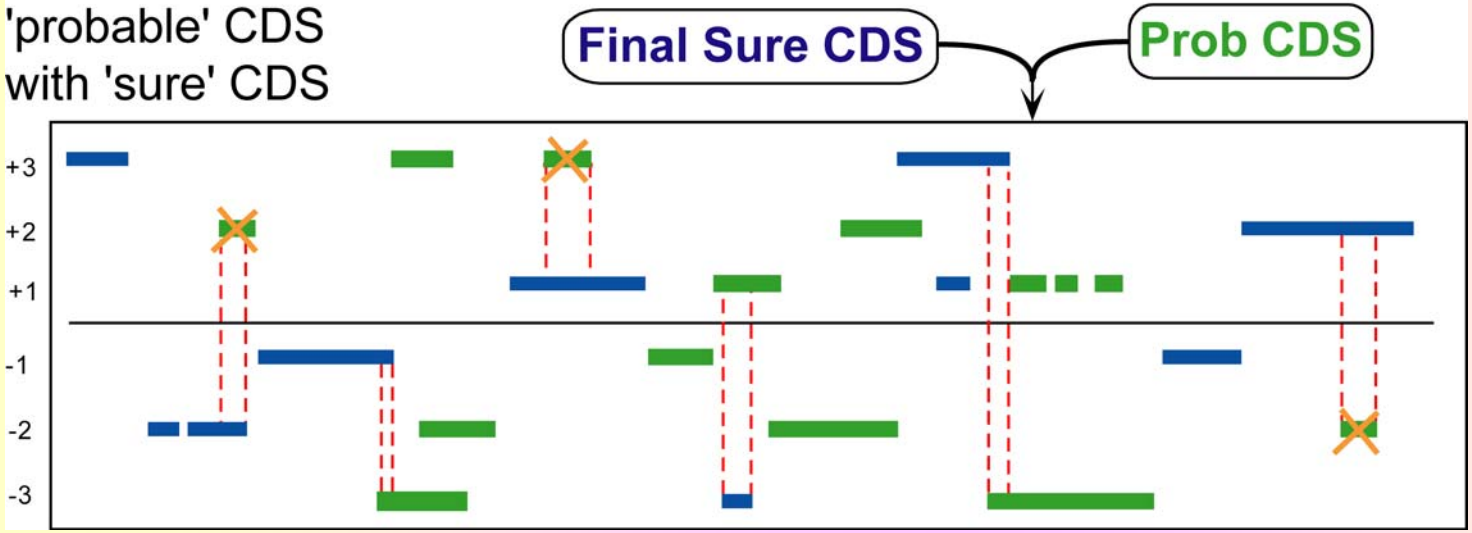
% Inclusion
par rapport à la longueur de la plus grande CDS.

Sure_ss_I = 10%

Sure_os_I = 30%

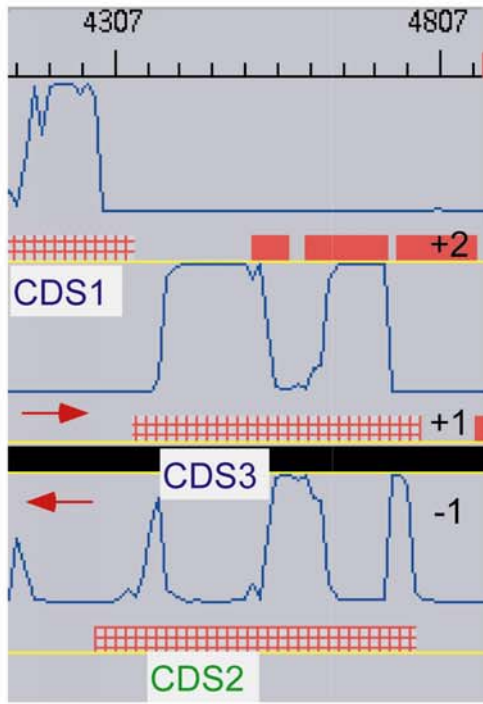


'probable' CDS
with 'sure' CDS



AMIFilter2

Bactérie	Nom	Début (brin)	L	Pc	%Chevauchement	Résultat
E. coli	CDS1	2795 (d)	1536	0.58		
	CDS2	4270 (r)	495	0.30	CDS1 / CDS2 12 CDS3 / CDS2 88	conservée éliminée
	CDS3	4330 (d)	444	0.58		



% Chevauchement
par rapport à la longueur
de la CDS probable

Sure_prob_os_O = 15%

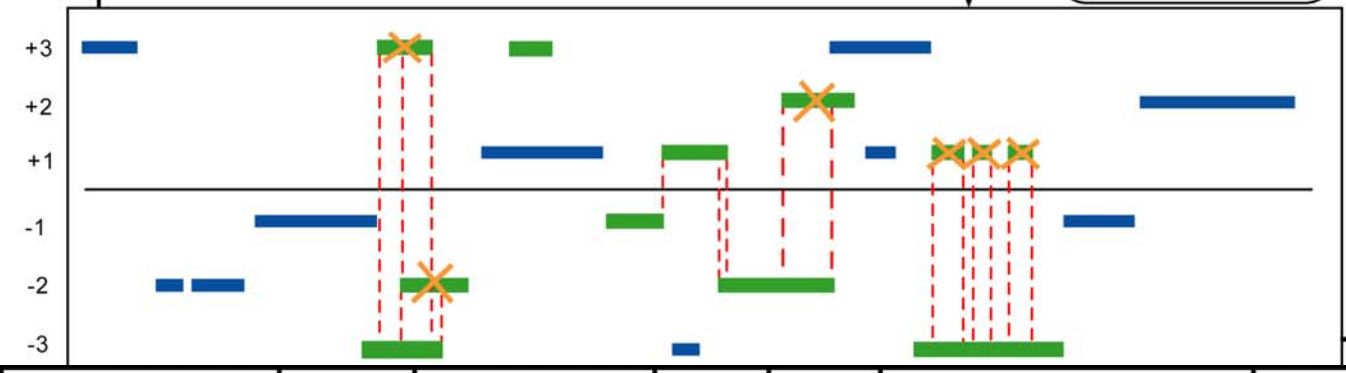
Chevauchement CDS même sens:
cas de frameshift OU
de réajustement du codon d'initiation

AMIFilter3

'probable' CDS
with 'probable' CDS

Final Sure CDS

Intermediate Prob CDS

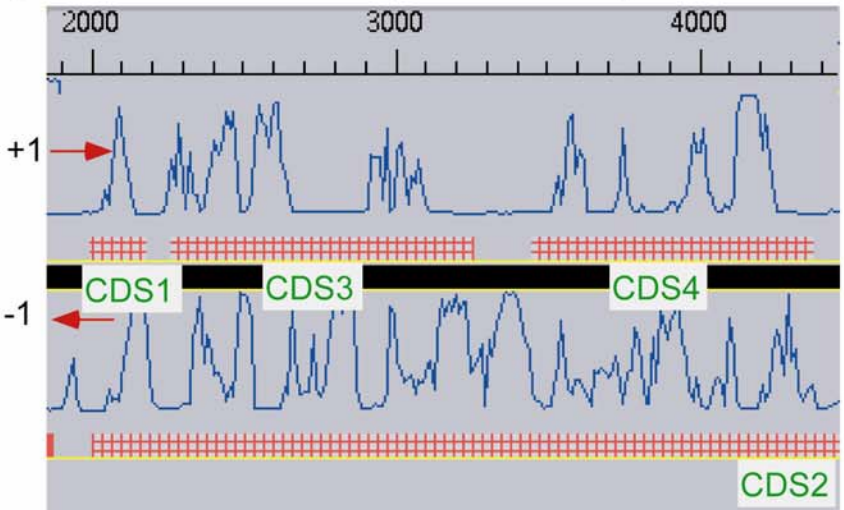


Most Likely CDS

Bactérie	Nom	Début (brin)	L	Pc	% Recouvrement	Score total	Résultat
M. tuberculosis	CDS1	1991 (d)	186	0.21	CDS2 / CDS1 96	96	éliminée
	CDS2	1997 (r)	3918	0.37	CDS1 / CDS2 4 CDS3 / CDS2 25 CDS4 / CDS2 24	53	conservée
	CDS3	2255 (d)	996	0.23	CDS2 / CDS3 100	100	éliminée
	CDS4	3437 (d)	933	0.23	CDS2 / CDS4 100	100	éliminée

score total de recouvrement
somme des %Chevauchement et/ou %Inclusion avec les autres CDS probables qui recouvrent la CDS courante

Prob_glob_IO = 80%



Prokaryotic Genome DataBase (PkGDB)

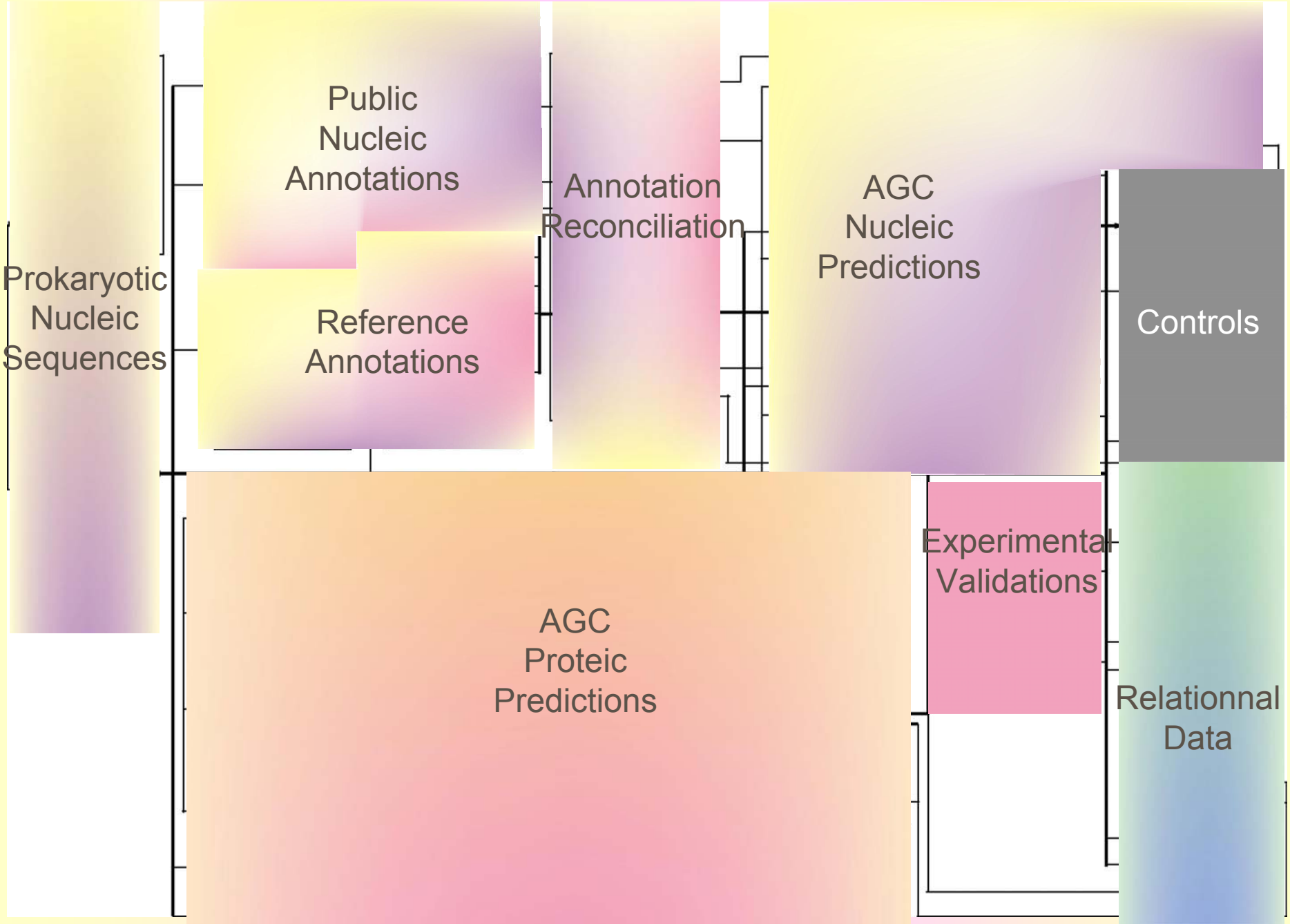
- **Structure** logique générique (PKGDB)
- Identification des **problèmes** d'annotation dans les **banques**
- SGBDR **MySQL**
- **Scripts** pour alimenter les tables
- **Instances** (PkGDB, NeisseriaDB, EnteroDB)

AGC dont S. Bocs

- Interface d'annotation et d'exploration **MaGe**

→ D. Vallenet

Prokaryotic Genome Database (PkGDB)



Processus de réannotation

- Stratégie de prédiction de CDS (**AMIMat / AMIGene**)
- Annotations des Banques (**RefSeq**)
- Base de données (**PkGDB**)
- Stratégie de **comparaison** de deux jeux de CDS d'un même chromosome
- Stratégies d'attribution de **statuts de similitude** et de **réannotation**

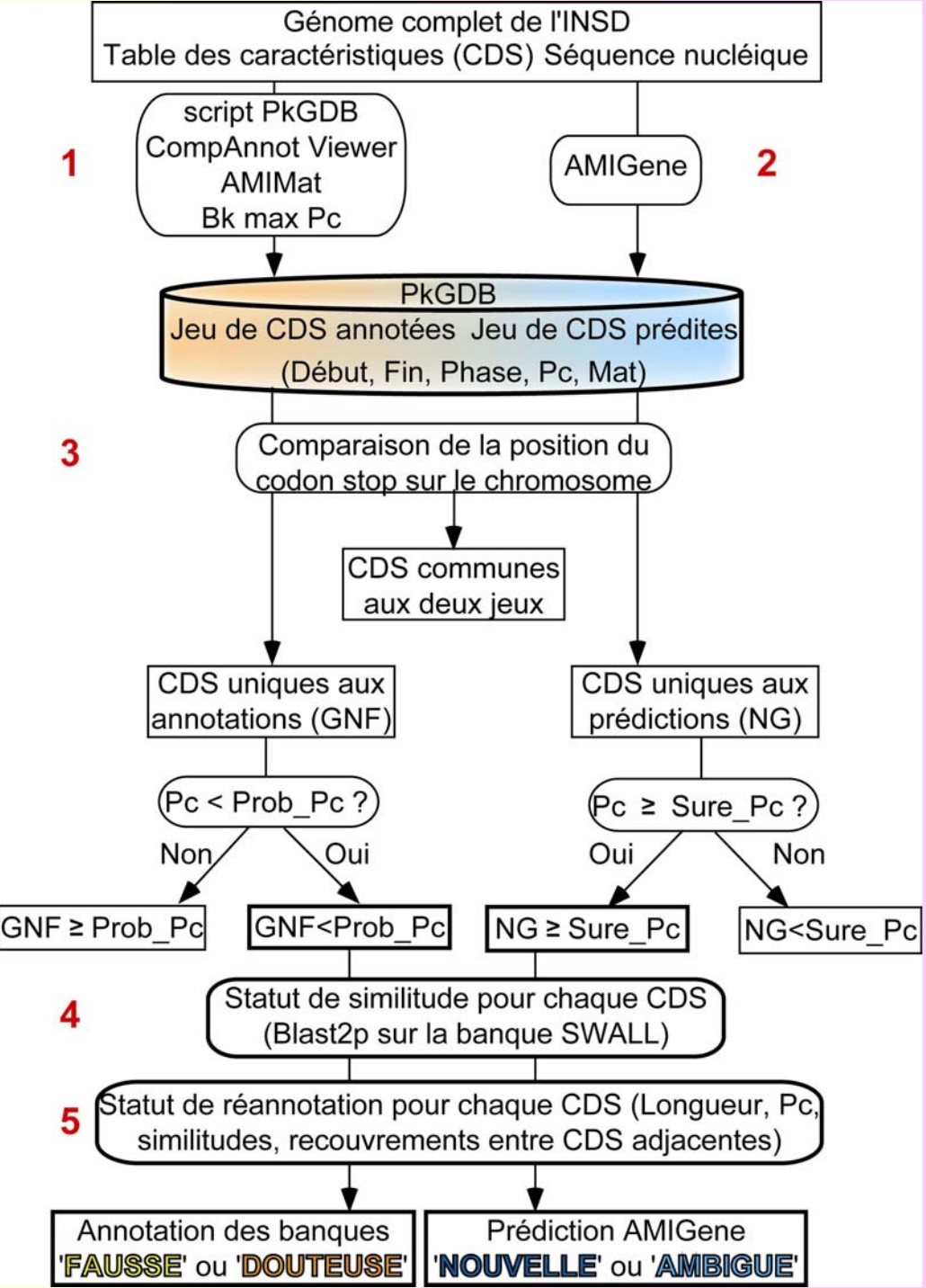
S. Bocs

- Application Web **Micheck** 

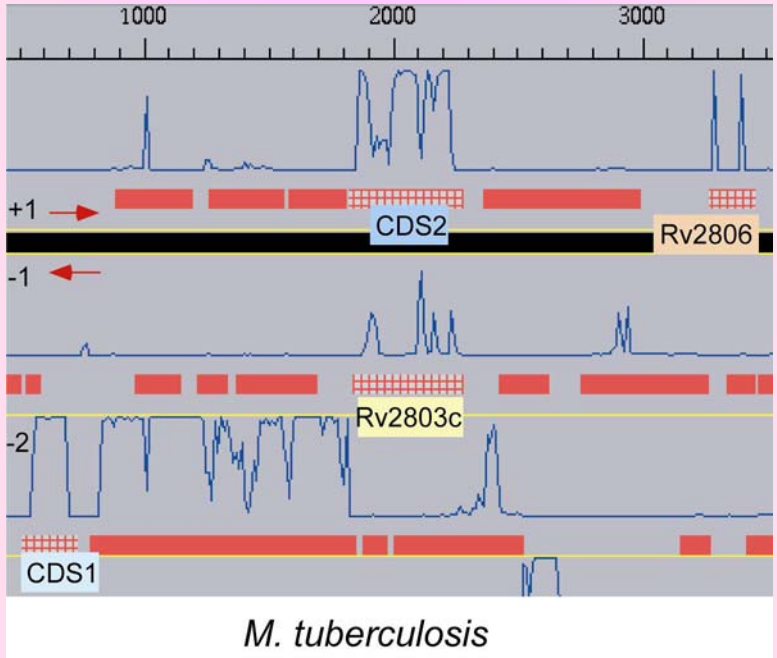
→ J. Le Saux

http://www.genoscope.cns.fr/agc/tools/micheck/

Processus de Réannotation



Exemple de statuts de réannotation



Polypeptide Requête					Meilleur polypeptide sujet					
Label	Début	B	L (pb)	Pc	Description	Espèce	L (aa)	E-value	% Identité	Statut
CDS1	3110507	R	228	0,620	NO SIMILARITY					ambigAGC
CDS2	3111819	D	465	0,560	HYPOTHETICAL 17.4 KDA PROTEIN	*M. tuberculosis*	158	1,5E-11	42	newAGC
Rv2803c	3111834	R	450	0,086	NO SIMILARITY					wrongBank
Rv2806	3113265	D	192	0,170	NO SIMILARITY					suspiBank

Tableau de résultats de réannotation

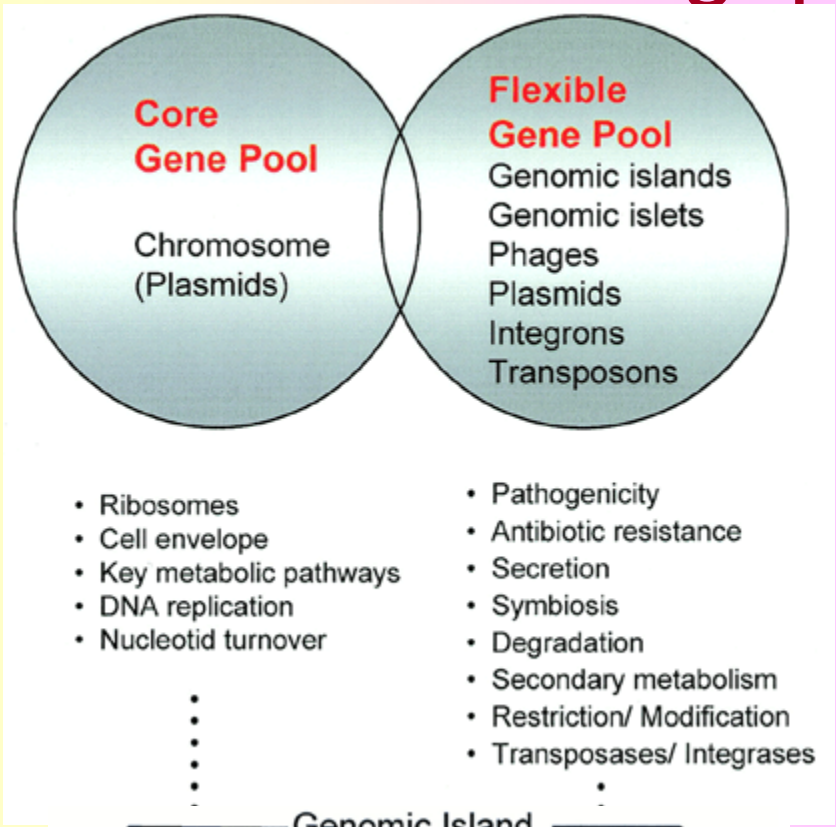
Species	Strain	Cl					Status				Status			
code		Nb	OA	AP	CC	(%)	W	(%)	S	(%)	N	(%)	A	(%)
ECOLI	K-12	3	4264	4553	4201	98,52	3	0,07	4	0,09	47	1,03	5	0,11
ECO57	EDL933	3	5339	5768	5181	97,04	14	0,26	39	0,73	42	0,73	12	0,21
YERPE	CO92	3	4108	4336	3981	96,91	2	0,05	70	1,70	13	0,30	14	0,32
YERPS	IP32953	3	3976	4282	3916	98,49	1	0,03	27	0,68	14	0,33	22	0,51
PHOLU	TT01	3	4905	5883	4839	98,65	0	0,00	24	0,49	17	0,29	51	0,87
NEIMA	Z2491	3	2147	2284	2069	96,37	6	0,28	44	2,05	16	0,70	39	1,71
NEIMB	MC58	3	2211	2383	2053	92,85	41	1,85	82	3,71	57	2,39	57	2,39
MYCTU	CDC1551	4	4273	4448	4035	94,43	45	1,05	125	2,93	59	1,33	26	0,58
MYCTU	H37Rv	4	3996	4387	3959	99,07	0	0,00	6	0,15	15	0,34	24	0,55
BACHD	C-125	3	4056	4470	3998	98,57	12	0,30	24	0,59	26	0,58	16	0,36
BACSU	168	3	4107	4586	4066	99,00	10	0,24	15	0,37	47	1,02	35	0,76

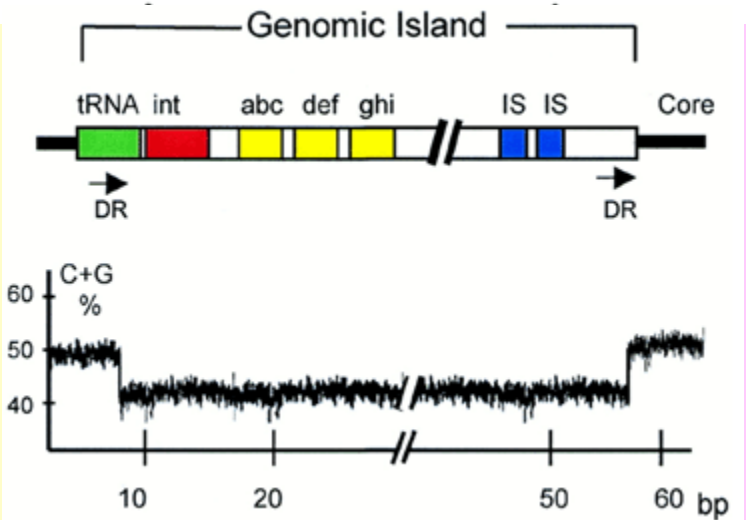
CDS, coding sequence, Nb Cl, class number; OA, original annotation; AP, AMIGA prediction; CC, common CDS to both OA and AP; GNF, gene not found; NG, new gene; Pc, coding average probability of a CDS; S, suspicious reannotation status; W, wrong; A, ambiguous; N, New

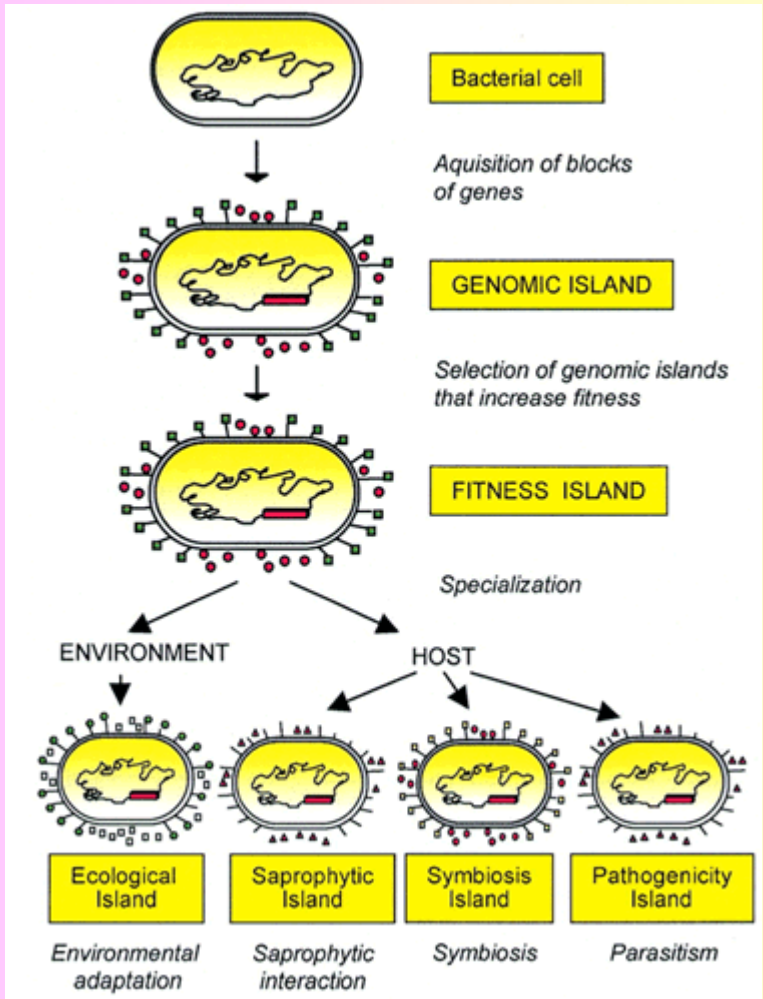
Applications biologiques de ces outils

- **Annotation** de nouveaux génomes:
  - *P. luminescens*
  - *Y. pseudotuberculosis*
- **Réannotation** de 30 génomes publics:
  - *B. subtilis*
  - *E. coli* K12
  - *M. tuberculosis*
- **Exploration** d'îlots génomique chez les entérobactéries:
  - *Y. pestis*
  - *P. luminescens*

Plasticité écologique et îlots génomiques

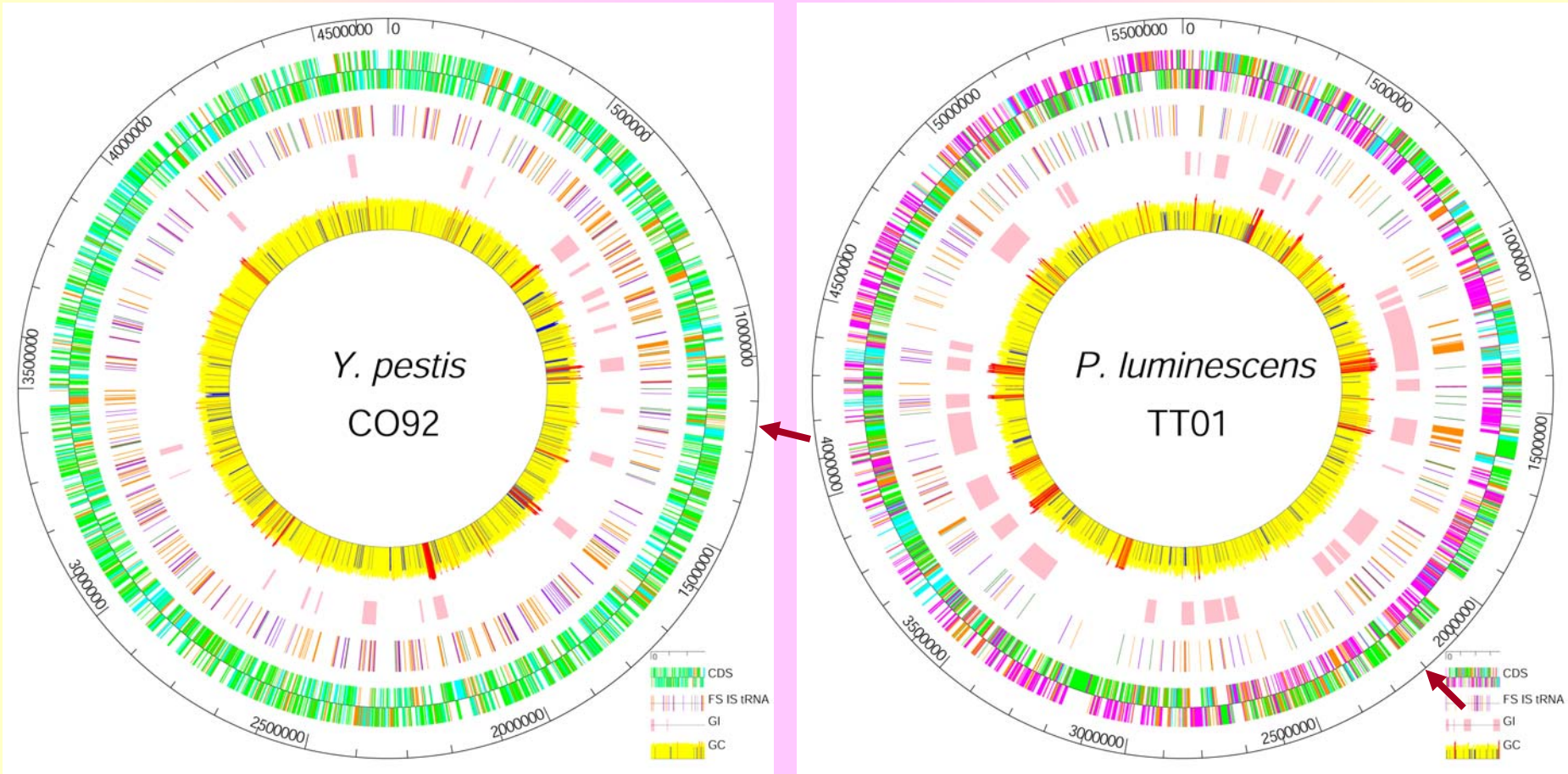




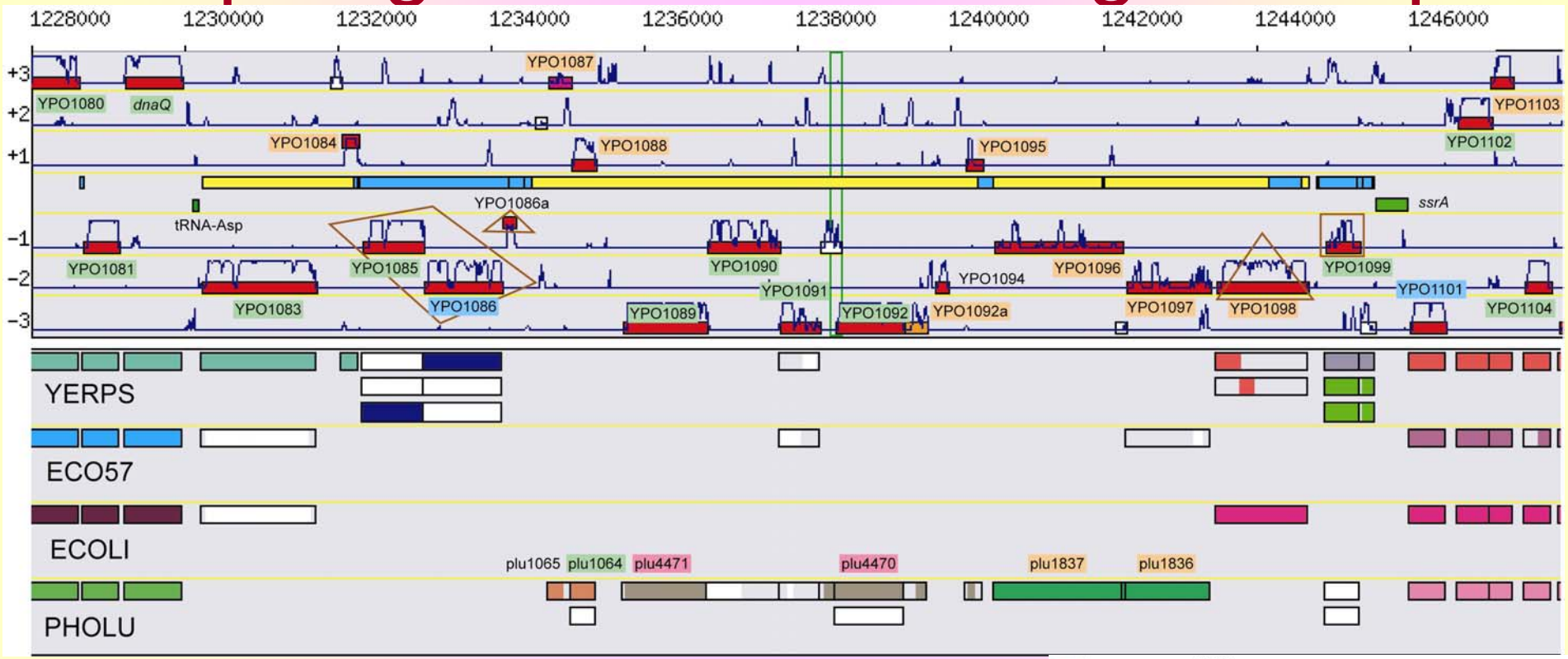


Hacker J. & Carniel E. (2001 Embo Report)

Vue globale d'îlots génomiques



Vue plongeante sur un îlot génomique



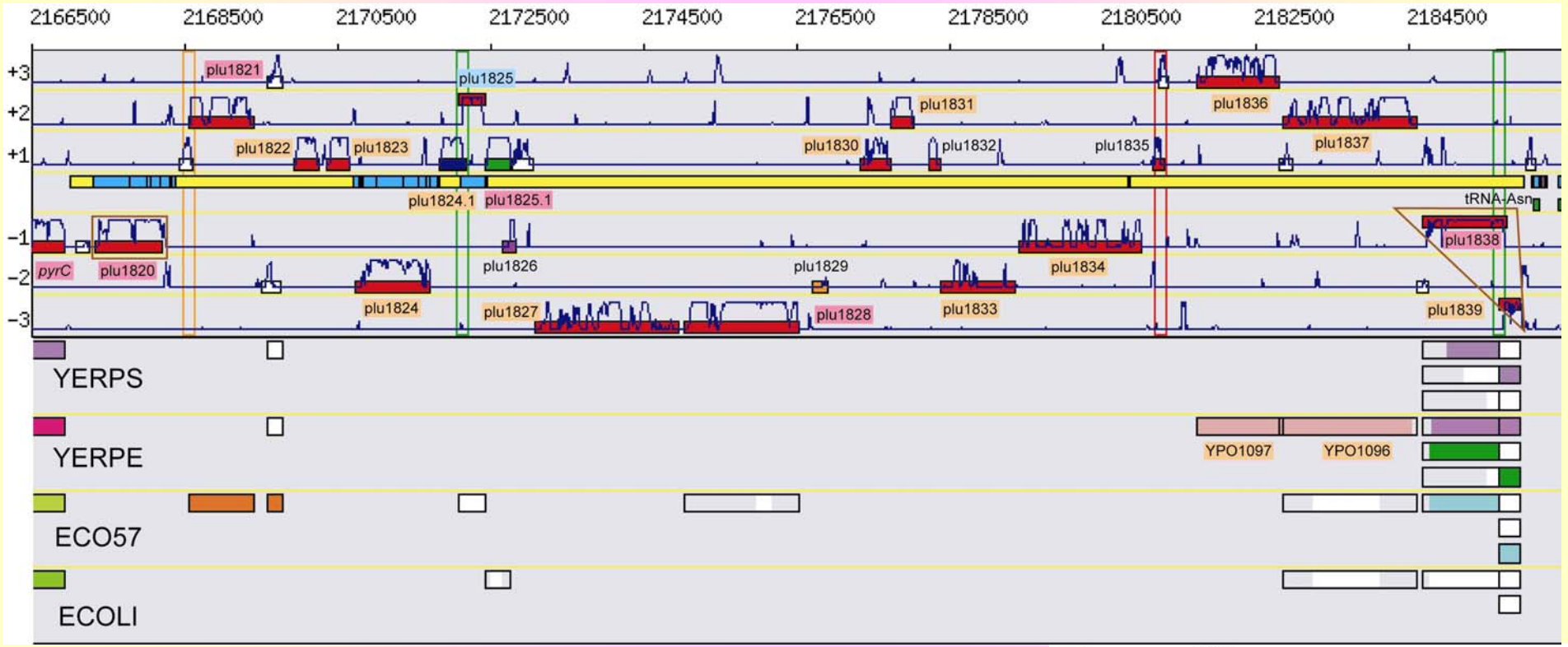
label	gene	type	Ψ	product
		tRNA	no	tRNA-Asp
YPO1083		CDS	no	putative permease
YPO1084		CDS	yes	CHP interrupted by IS100
YPO1085	ypmt1.57c	CDS	no	IS100, ATP-binding protein
YPO1086	y1093	CDS	no	transposase for IS100
YPO1086a	intA	CDS	yes	phage integrase (partial)
YPO1087		CDS	no	putative prophage protein
YPO1088		CDS	no	putative DNA-binding prophage protein
YPO1089		CDS	no	putative regulatory prophage protein
YPO1090		CDS	no	putative prophage DNA primase
YPO1091		CDS	no	putative prophage protein
YPO1092		CDS	no	putative DNA-binding prophage protein

label	gene	type	Ψ	product
YPO1092a		CDS	no	HP
YPO1094		CDS	no	HP
YPO1095		CDS	no	HP
YPO1096		CDS	no	putative phage protein
YPO1097		CDS	no	putative phage protein
YPO1098		CDS	no	putative prophage integrase
YPO1099	tnp	CDS	no	transposase for the IS1541
	ssrA	RNA	no	tmRNA [10Sa RNA]
YPO1101	smpB	CDS	no	SsrA-binding protein (small protein B)
YPO1102		CDS	no	CHP
YPO1103		CDS	no	CHP
YPO1104		CDS	no	CHP
YPO1105		CDS	no	DNA repair protein RecN

common CDS
uniqBank
noStatusBank
suspiciousBank
wrongBank
newAGC
ambiguousAGC
noStatusAGC
uniqAGC
YPO1080 Class I
YPO1101 Class II
YPO1087 Class III
plu4471 Class IV
Phage integrase
IS transposase

ProFED Frameshift
weak
medium
strong
Annotated Frameshift
Genomic Objet without frame
Genomic island
Repetition
tRNA

Vue plongeante sur un îlot génomique



label	gene	type	Ψ	product
plu1820	ISPlu6G	CDS	no	Transposase, IS982 family
plu1821		CDS	no	CHP
plu1822		CDS	no	HP
plu1823		CDS	no	HP
plu1824	ISPlu1D	CDS	no	Transposase, IS30 family
plu1824.1		CDS	no	HP
plu1825		CDS	yes	Truncated transposase, IS4 family
plu1825.1		CDS	no	CHP
plu1826		CDS	no	Hypothetical gene
plu1827		CDS	no	CHP
plu1828		CDS	no	CHP
plu1829		CDS	no	Hypothetical gene

label	type	Ψ	product
plu1830	CDS	no	HP
plu1831	CDS	no	HP
plu1832	CDS	no	Hypothetical gene
plu1833	CDS	no	CHP
plu1834	CDS	no	CHP
plu1835	CDS	no	Hypothetical gene
plu1836	CDS	no	putative phage protein
plu1837	CDS	no	putative phage protein
plu1838	CDS	yes	C-term of putative integrase
plu1839	CDS	yes	N-term of putative integrase
	tRNA	no	transfert RNA-Asn

common CDS
uniqBank
noStatusBank
suspiciousBank
wrongBank
newAGC
ambiguousAGC
noStatusAGC
uniqAGC
YPO1080 Class I
YPO1101 Class II
YPO1087 Class III
plu4471 Class IV
Phage integrase
IS transposase

ProFED Frameshift
weak
medium
strong
Annotated Frameshift
Genomic Objet without frame
Genomic island
Repetition
tRNA

Conclusion sur l'exploration d'îlots génomiques

- FS
- Gènes de mobilités (IS, transposase, intégrase)
- tRNA
- Répétitions
- Région atypique (% G3+C3, classe AFC)
- Mots clés dans la fonction protéique des gènes (virulence, résistance, phage)

Conclusions / Perspectives

- **Améliorations** du processus global de (Ré)annotation
  - PkGDB/MaGe (scripts, structure, formats des banques)
  - AMIMat/AMIGene (HMM, classification, start)
  - Attribution d'un statut de réannotation/Micheck (seuillage)
- **Application** de ce processus à
  - des génomes publics (HAMAP)
  - de nouveaux génomes (Genoscope)
- **Développement** d'une stratégie d'îlots génomique afin d'élucider les clés du transfert horizontal dans l'évolution de la microbiodiversité