Transferts horizontaux de gènes chez les bactéries

Guy Perrière

Pôle Bioinformatique Lyonnais
Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558

perriere@biomserv.univ-lyon1.fr

Définitions

- Transfert vertical :
  - Transmission de l'information génétique de la génération parentale à la progéniture.
- Transfert horizontal (ou latéral) :
  - Passage de séquences d'un génome à un autre.
  - Mise en jeu de mécanismes comme la transformation, la transduction et la conjugaison.
  - Intégration par recombinaison.

Étude des transferts

- Les transferts horizontaux semblent toucher l'ensemble des organismes vivants.
- La plupart des études ont été réalisées chez les bactéries :
  - Raisons historiques (Griffith, 1928).
  - Motivations d'ordre médical et économique.
  - Rareté des transferts procaryotes ➔ eucaryotes :
    - Nécessité d'intégrer les lignées germinales.

Prérequis

- Proximité entre le donneur et l'accepteur.
- Stabilité de l'ADN dans l'environnement.
- Utilisation d'un vecteur de transmission.
- Capture par l'hôte suivi d'une insertion.
- Maintient de l'ADN incorporé.
- Sélection.

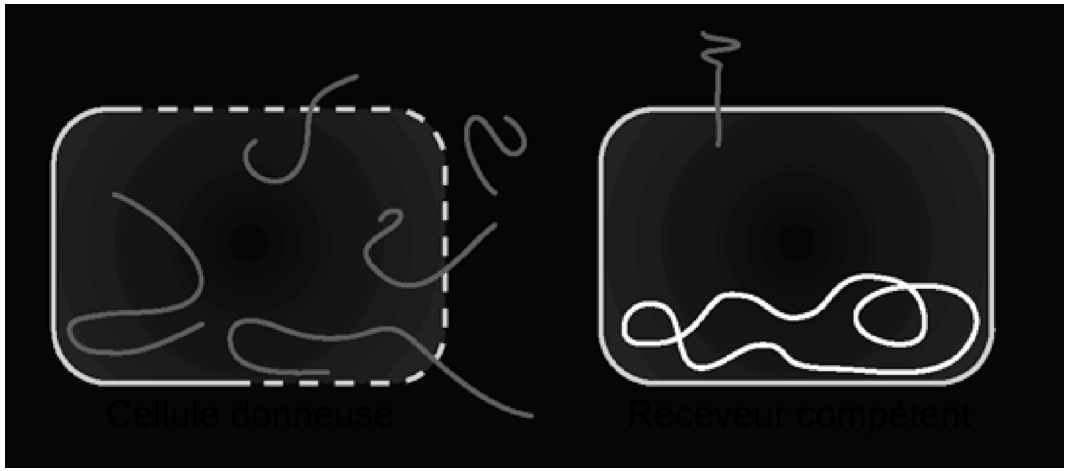
Protections / limitations

- Instabilité de l'ADN internalisé.
- Systèmes de restriction / modification.
- Incompatibilité compositionnelle (G+C, composition en codons).
- Pas de régions recombinantes suffisamment longues.

Cas connus

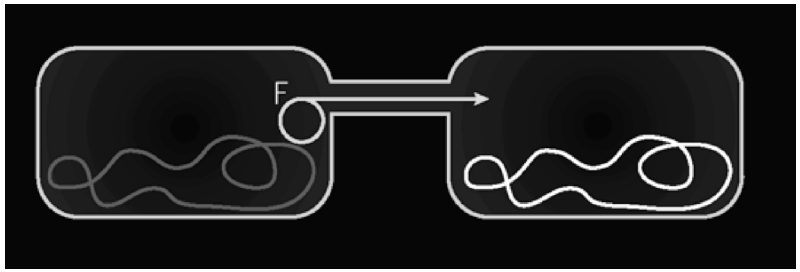
- Gènes plasmidiques de résistance aux anti-biotiques et aux toxines.
- Séquences d'insertion (IS).
- Ilôts de pathogénicité.
- Plasmide Ti d'*Agrobacterium tumefasciens*.
- Phages et pseudo-phages.
- Transfert depuis des génomes d'organites (mitochondrie, chloroplaste) vers le noyau.

La transformation

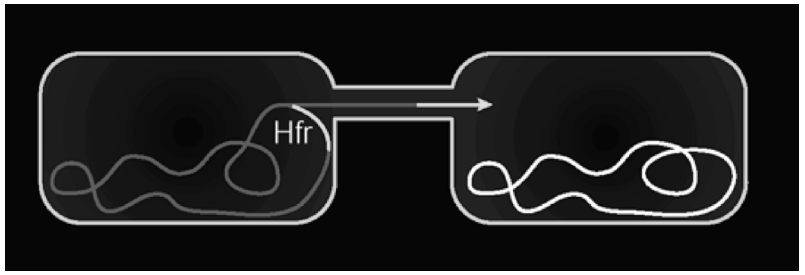


- Capture par la cellule de fragments d'ADN nu présents dans le milieu.
- La transformation naturelle est restreinte à certaines espèces bactériennes « compétentes ».

La conjugaison



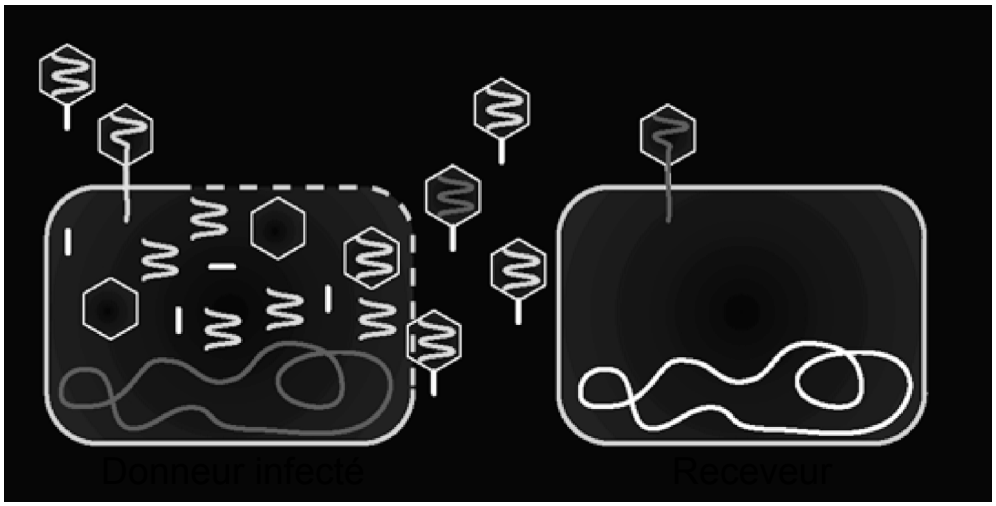
Évènement fréquent



Évènement rare

- Transfert par l'intermédiaire d'un plasmide conjugatif (*e.g.*, le plasmide *F* chez *E. coli*).
- La conjugaison serait le mécanisme le plus courant d'échange d'ADN entre organismes distants.

La transduction

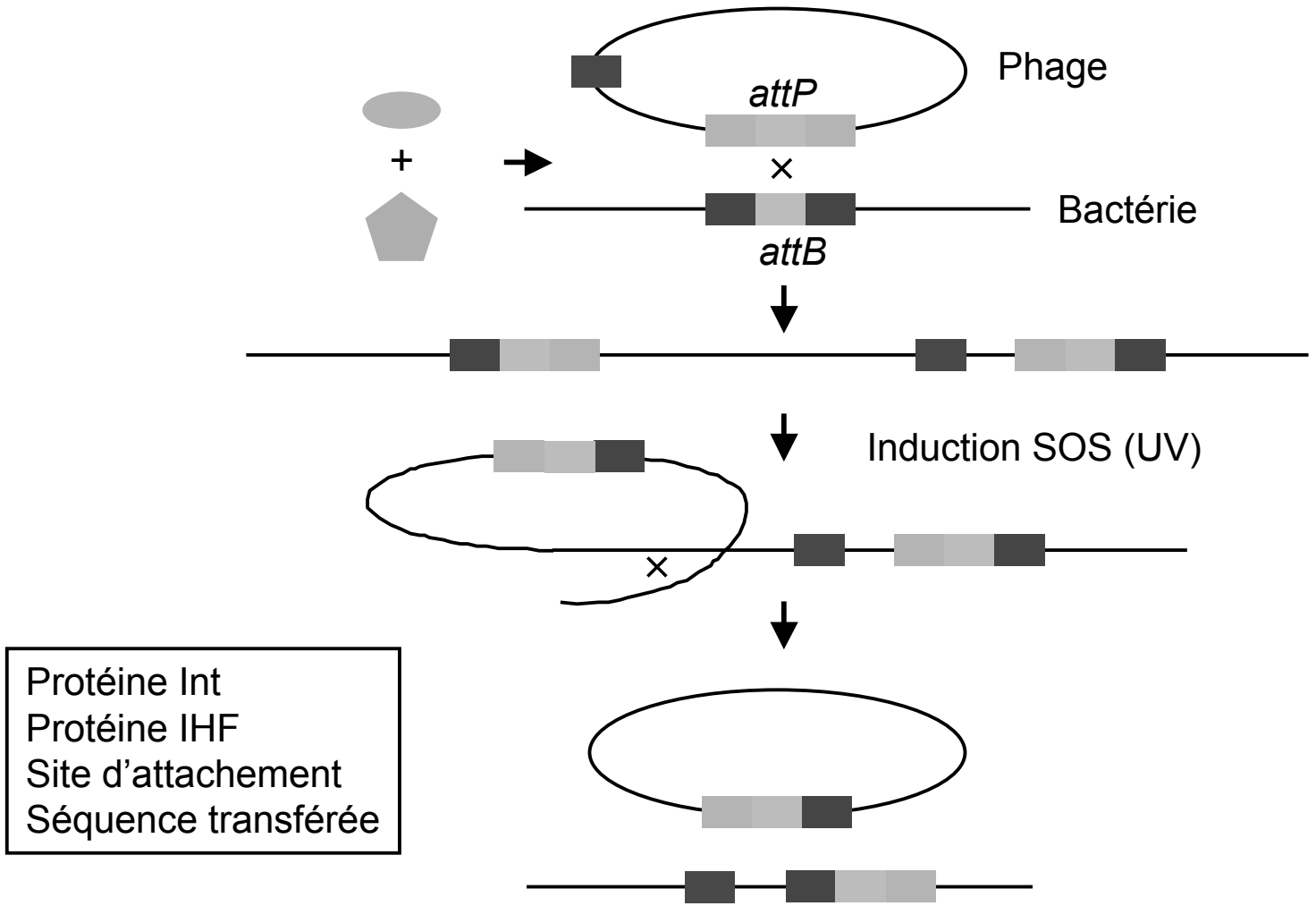


- Transfert d'ADN par l'intermédiaire de bactério-phages.
- Implique généralement des bactéries appartenant à des espèces proches (spécificité d'hôte) :
  - Certains phages (*e.g.*, μ) possèdent un large spectre.

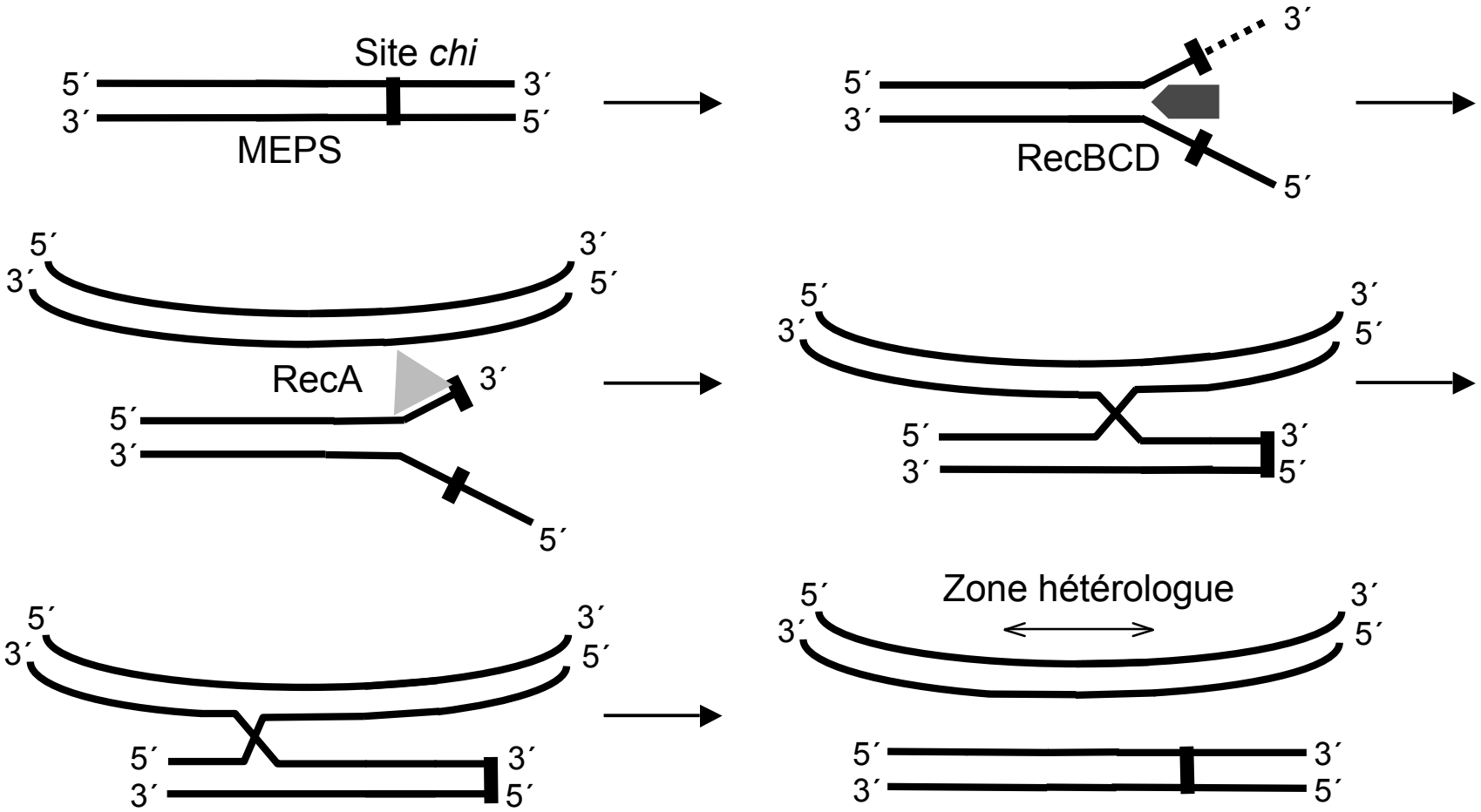
Intégration des séquences

- L'intégration des fragments exogènes est rendue possible par différents mécanismes :
  - Recombinaison site-spécifique.
  - Recombinaison générale homologue :
    - Recombinaison réciproque.
    - Recombinaison non réciproque.
- Contribution à la divergence des génomes :
  - Chez *E. coli*, effet 50 fois plus important que la mutagenèse.

Recombinaison site-spécifique



Recombinaison réciproque



Les zones MEPS

- Zones homologues minimales nécessaire à l'amorçage de la recombinaison (*Minimum Efficient Processing Segment*) :
  - Doivent présenter une forte similarité entre l'ADN endogène et exogène.
  - Longueur dépendante des mécanismes de réparation des mésappariements existant chez l'hôte :
    - *E. coli* : 23-27 pb.
    - *B. subtilis* : ≈ 70 pb.

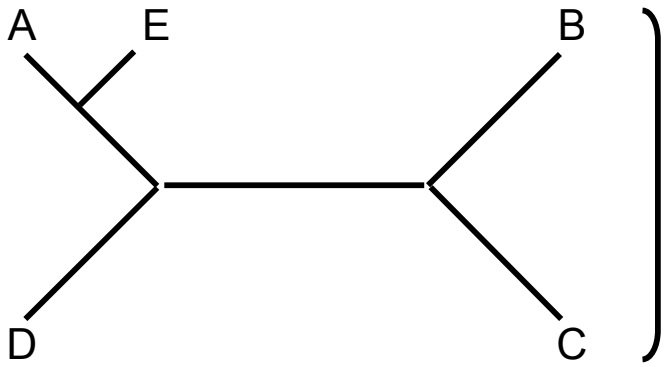
Rôle des site *chi*

- Séquences d'ADN courtes (10 pb) stimulant la recombinaison.
- Permettent le démarrage de la recherche des zones MEPS par la protéine RecA.
- Une région comprenant une zone MEPS et un site *chi* peut recombiner.
- Leur présence peut permettre de confirmer l'existence d'un transfert horizontal.
- Connus dans un petit nombre d'espèces.

Méthodes extrinsèques

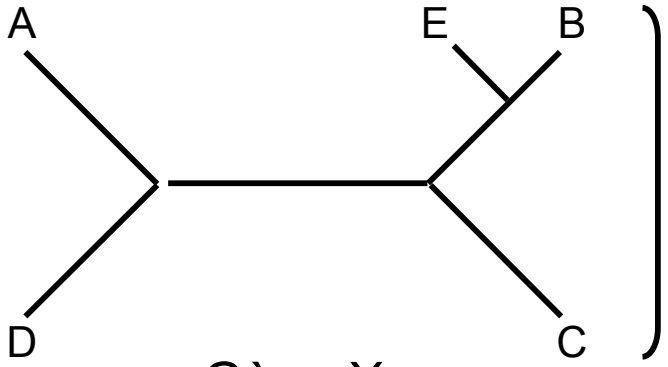
- Se basent sur des données de la phylogénie moléculaire :
  - Réalisables sur des gènes pour lesquels on trouve un homologue dans plusieurs espèces.
  - Problèmes :
    - Nécessitent l'utilisation de gènes orthologues.
    - Limites techniques des méthodes de reconstruction.
    - Nécessité d'effectuer les comparaisons par rapport à un arbre de référence :
      - ✓ Quel arbre choisir ?

Phylogénies et transferts



Référence

Phylogénie de référence
pour A, B, C, D et E



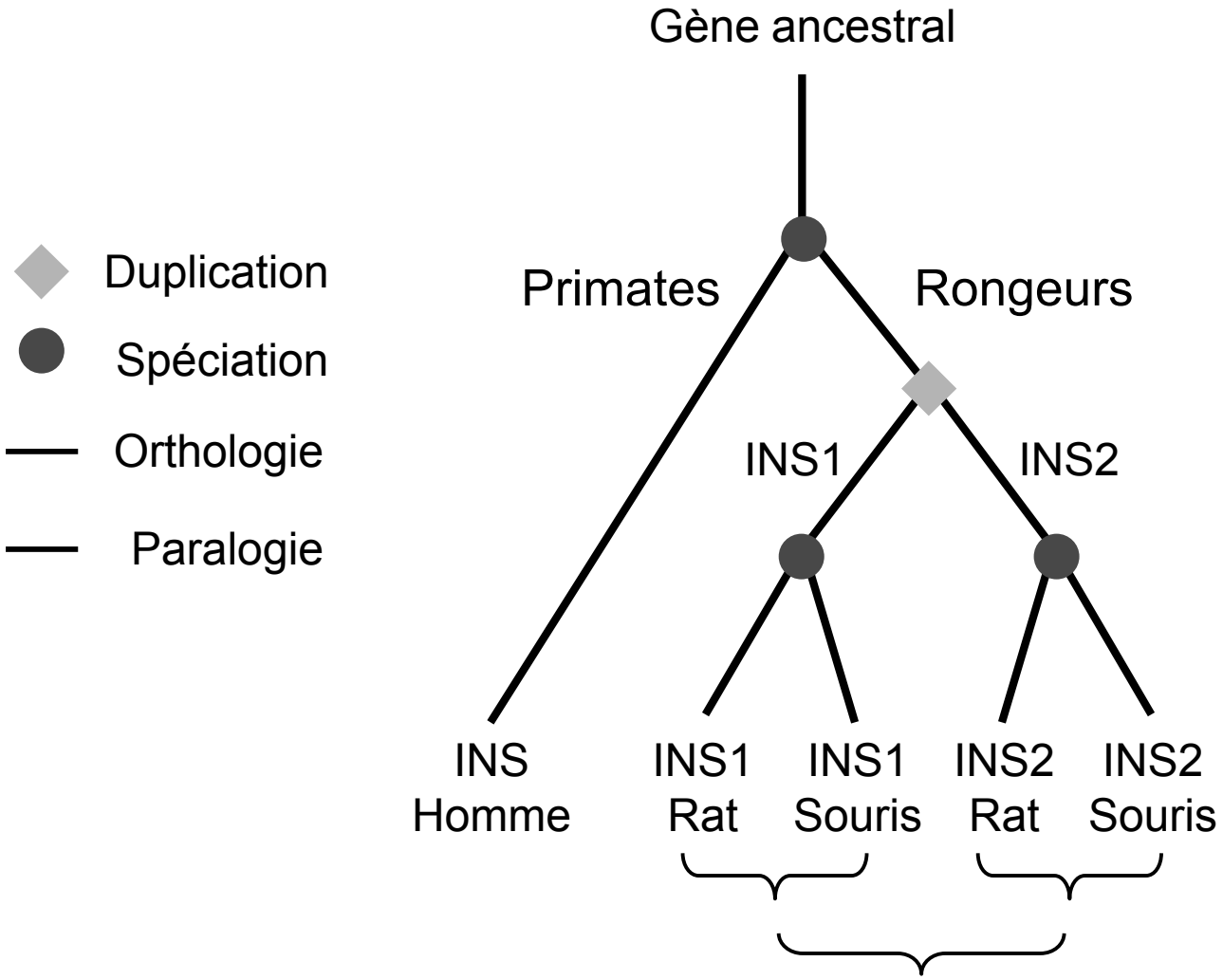
Gène X

Transfert horizontal de
B → E pour le gène X

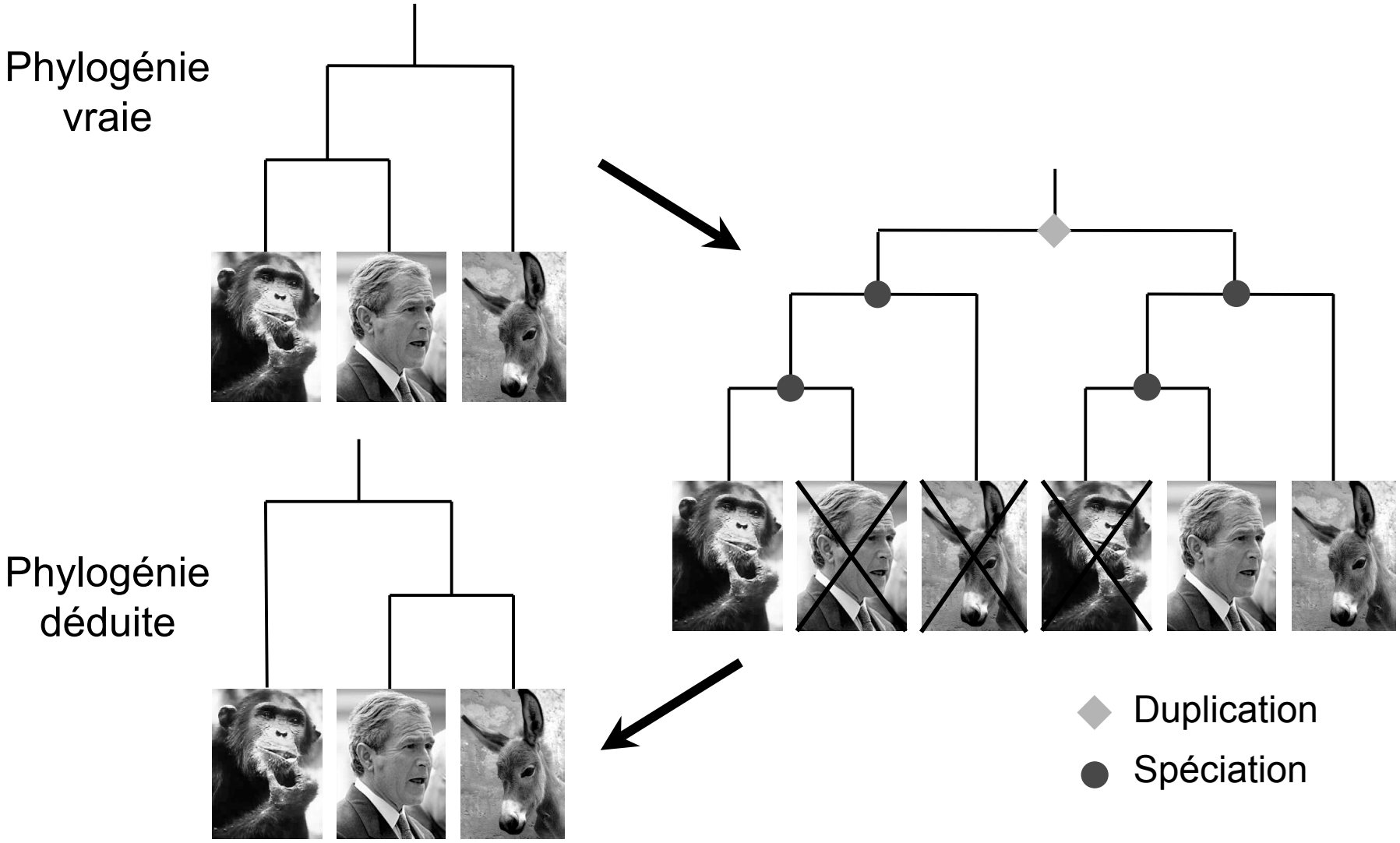
Notion de gènes homologues

- La phylogénie moléculaire se base sur l'utilisation de gènes homologues :
  - Deux séquences sont dites homologues si elles possèdent un ancêtre commun.
  - L'existence d'un ancêtre commun est inférée à partir de la similarité.
  - Seuil pour les protéines :
    - 30 % d'identité sur une longueur de 100 AA $\Rightarrow$ homologie entre les séquences.

Orthologues et paralogues



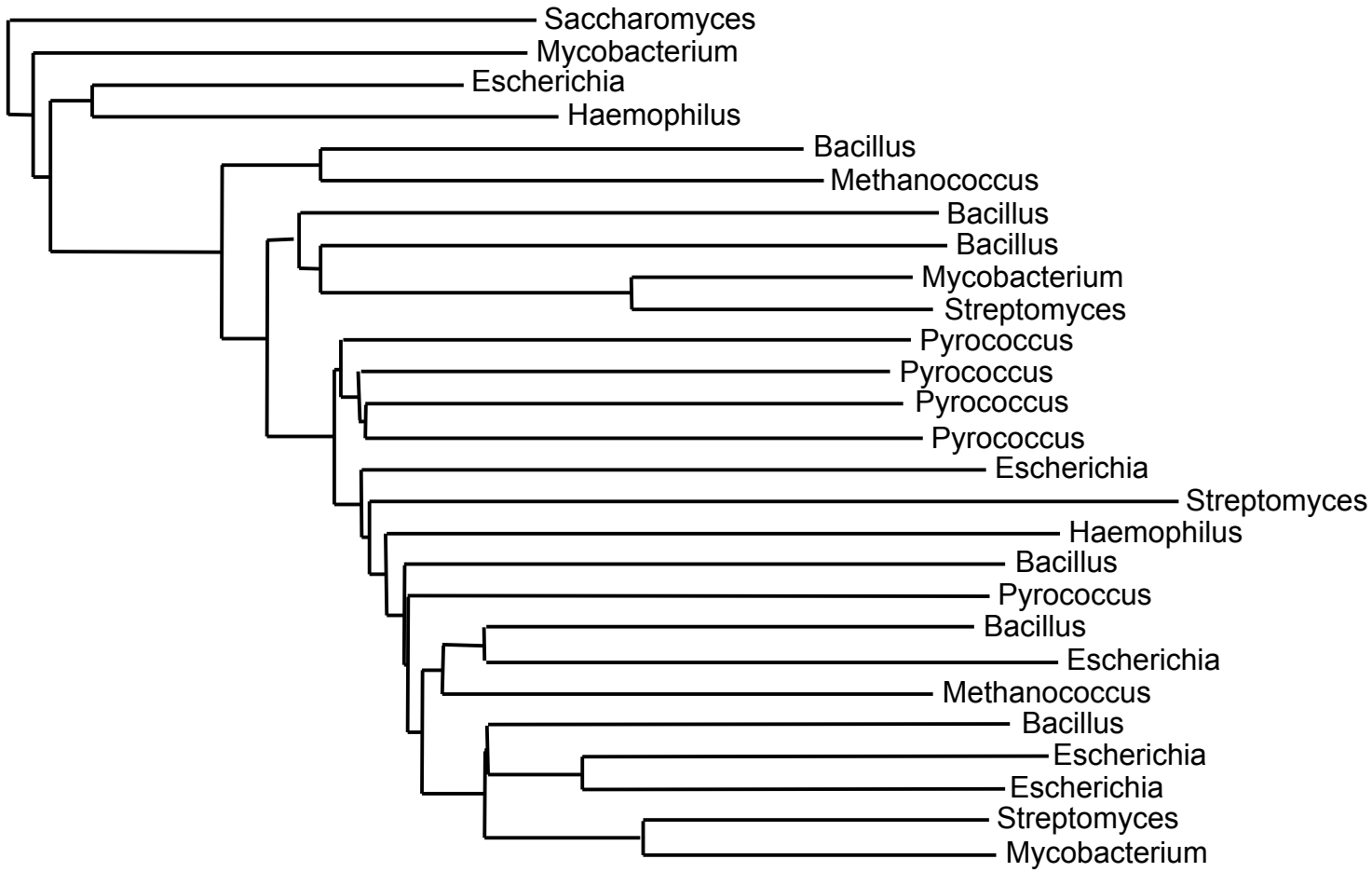
Paralogues et phylogénies



Les paralogues sont fréquents

- 30 % des gènes chez *E. coli* (compte non tenu des possibles paralogies cachées).
- Existence de duplications multiples :
  - Les relations d'orthologie sont souvent non bijectives.
- Divergences pouvant être importantes après duplication :
  - Difficulté d'identifier de nombreux paralogues.

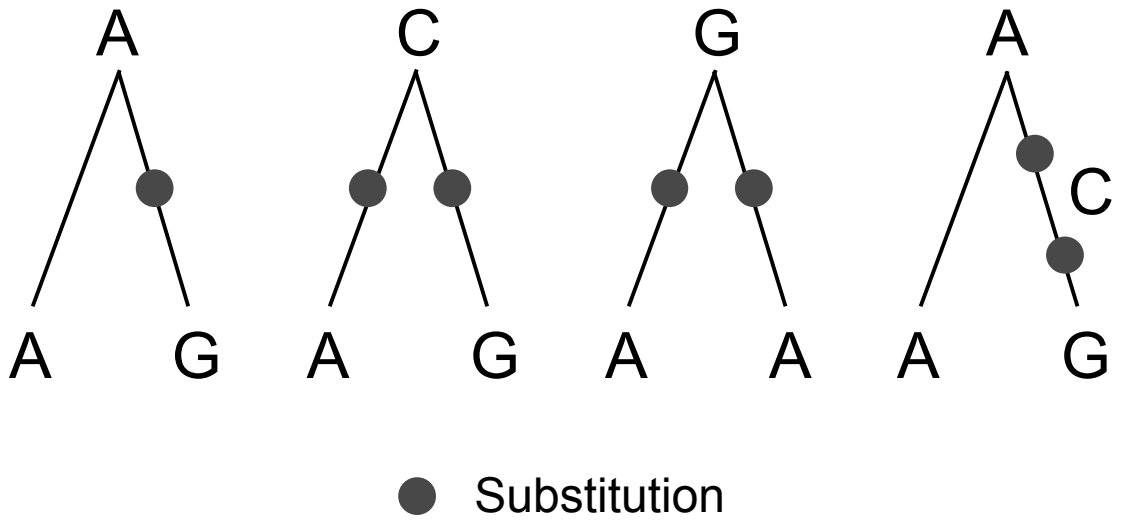
Exemple



Aminotransférases pyridoxal-phosphate dépendantes (III)

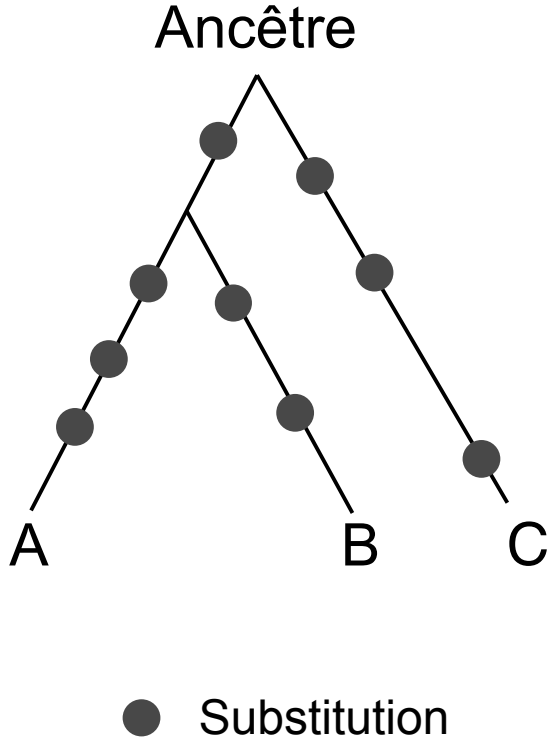
Substitutions multiples

- La distance évolutive réelle (*D*) est généralement supérieure aux différences observées (*d*).
- En faisant des hypothèses sur la nature du processus évolutif, on peut estimer *D* à partir de *d*.

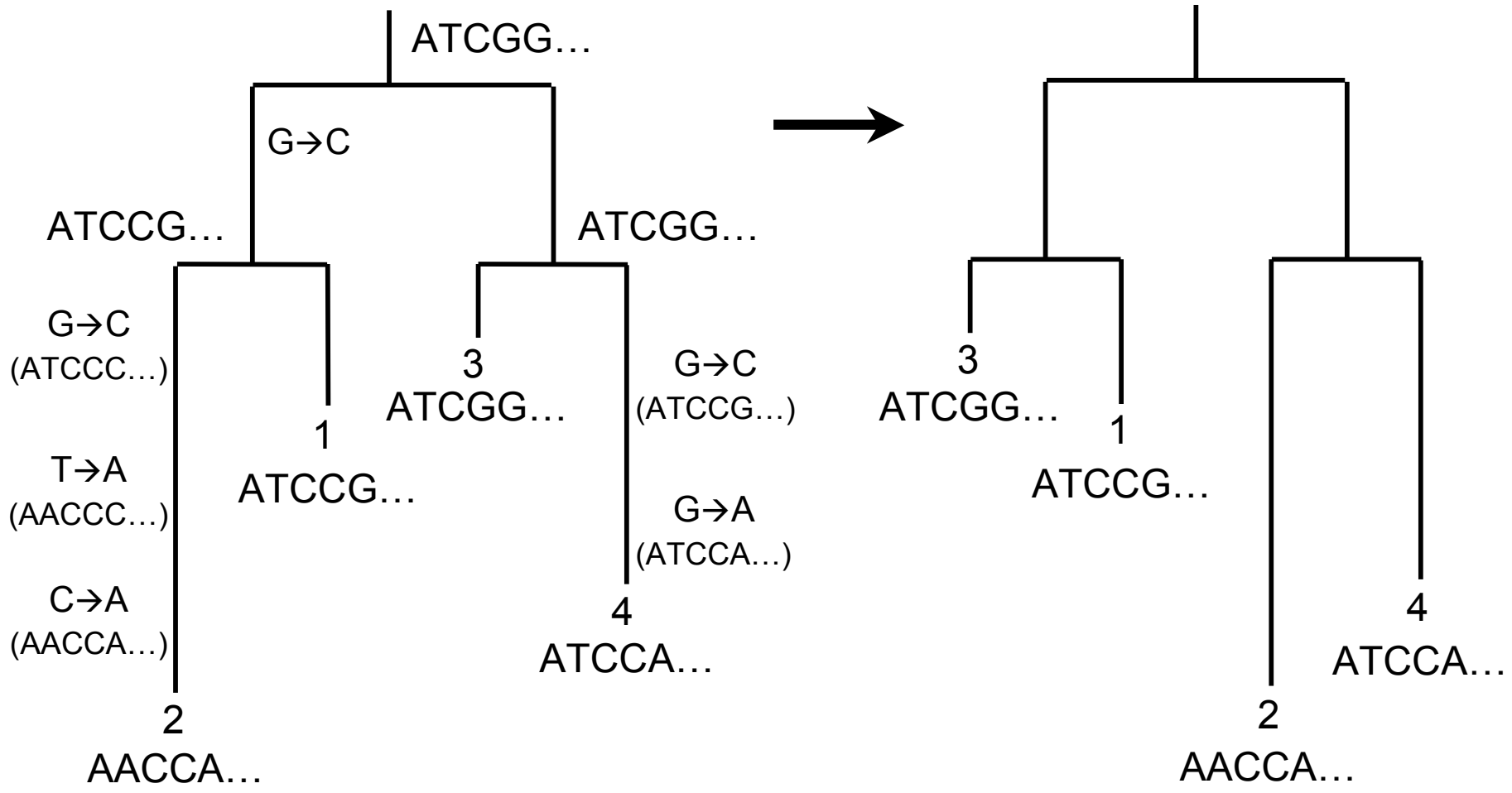


Phénomène de saturation

- Trop de substitutions depuis la divergence avec l'ancêtre commun :
  - Perte du signal phylogénétique.
  - Impossibilité de reconstruire l'arbre vrai quelle que soit la méthode employée.



Artefact des longues branches

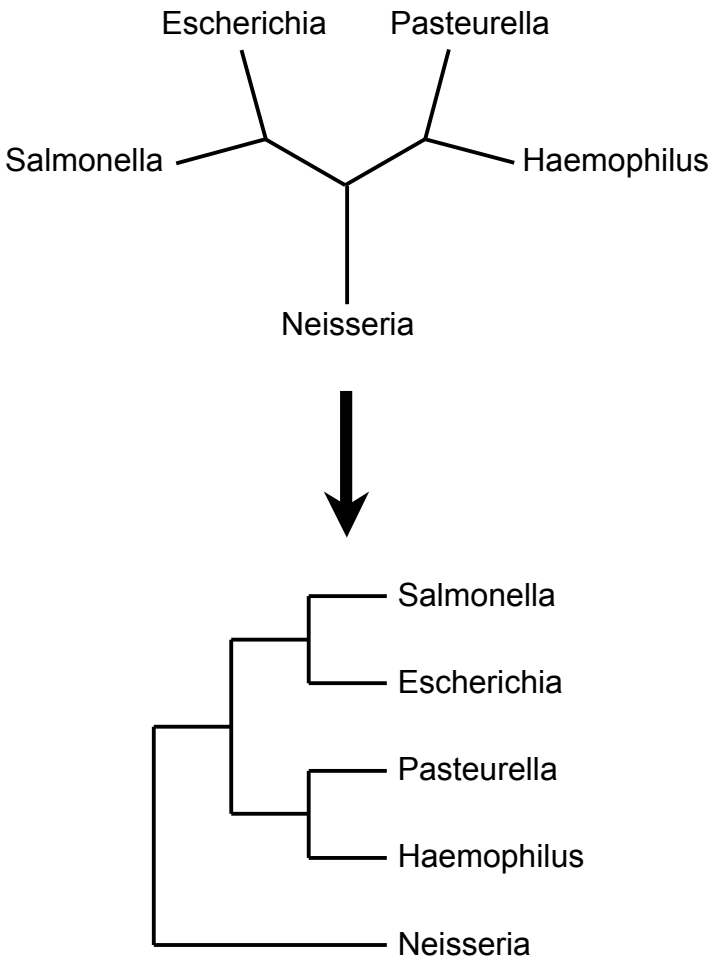


Arbre vrai

Arbre observé

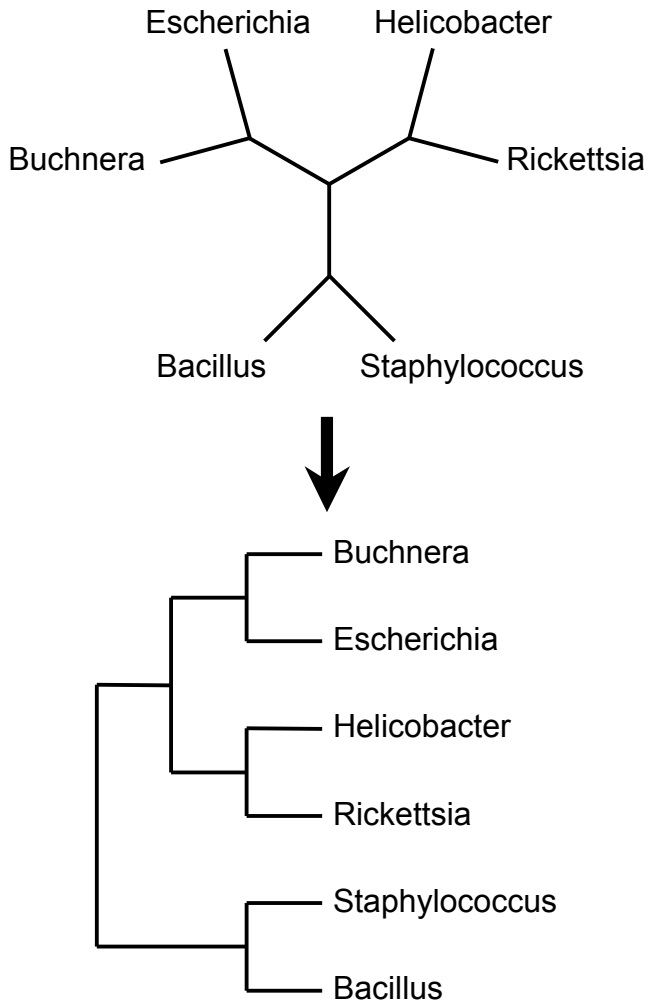
Racinement par un *outgroup*

- Choix du groupe :
  - Une espèce ou un groupe d'espèces.
  - Ni trop proche ni trop éloigné :
    - Degré de divergence entre les organismes considérés.
  - Exemple 1 :
    - γ-Protéobactéries racinées avec une β-Protéobactérie.



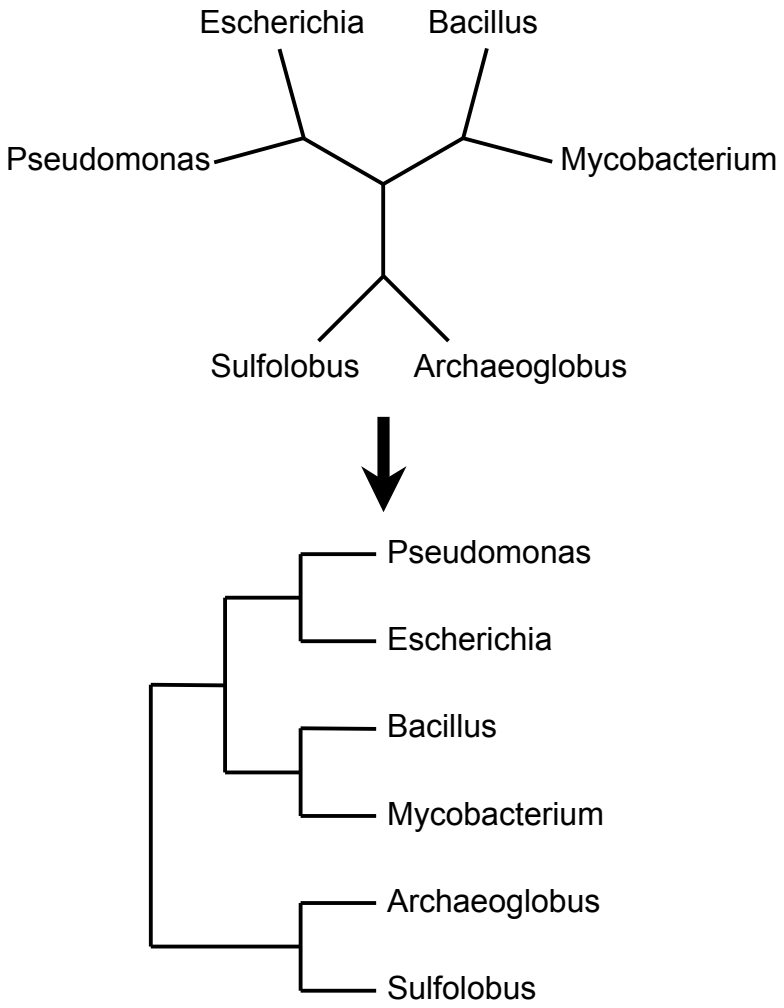
Racinement par un *outgroup*

- Choix du groupe :
  - Une espèce ou un groupe d'espèces.
  - Ni trop proche ni trop éloigné :
    - Degré de divergence entre les organismes considérés.
  - Exemple 2 :
    - Protéoactéries racinées avec des Firmicutes.



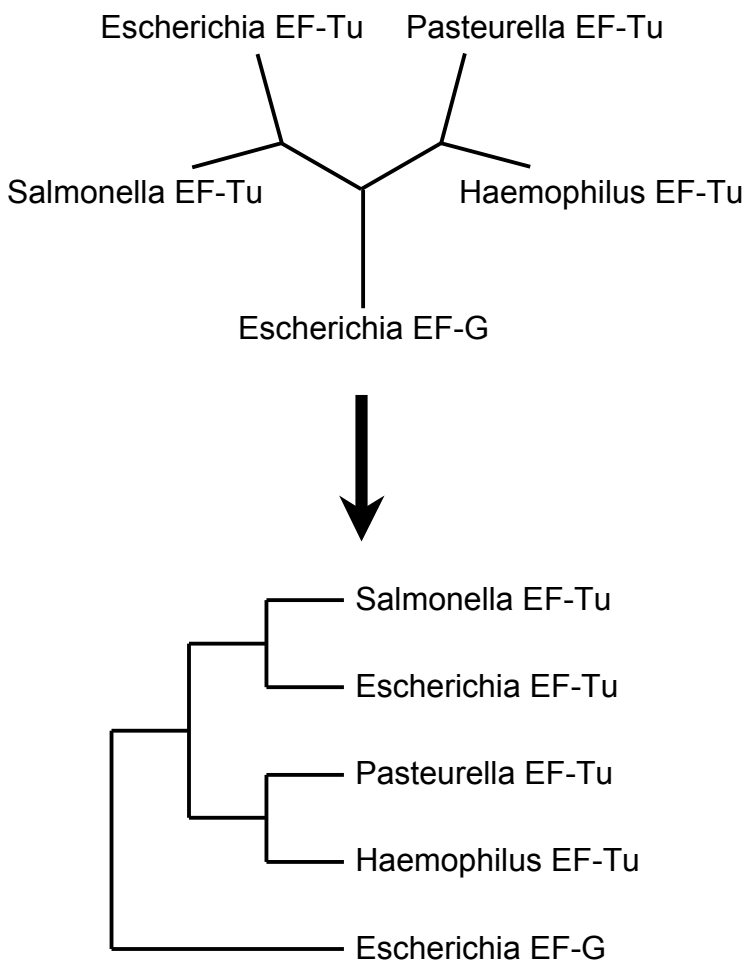
Racinement par un *outgroup*

- Choix du groupe :
  - Une espèce ou un groupe d'espèces.
  - Ni trop proche ni trop éloigné :
    - Degré de divergence entre les organismes considérés.
  - Exemple 3 :
    - Bactéries racinées avec des Archées.

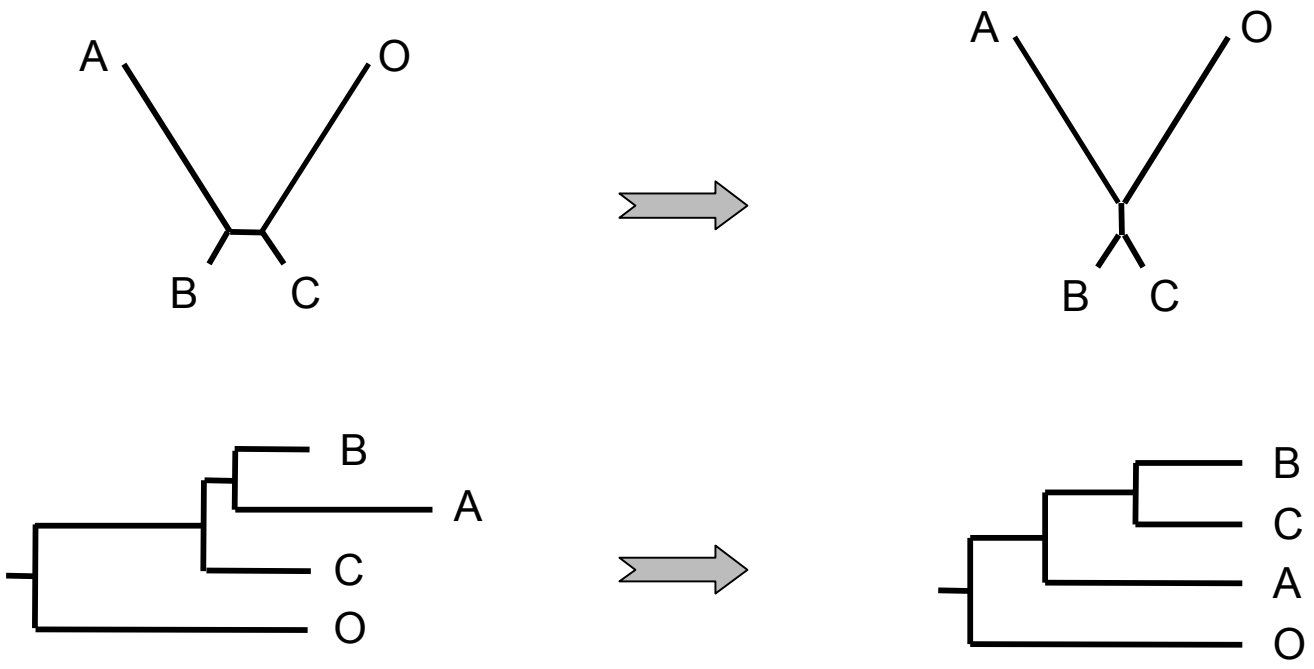


Racinement par un paralogue

- Duplication chez l'ancêtre commun des espèces étudiées :
  - Racinement en utilisant une des copies paralogues.
  - Exemple :
    - Facteurs d'élongation de la traduction EF-Tu/1α et EF-G/2 (Bactéries/ Archées et Eucaryotes).



Effet de l'*outgroup*



Arbre vrai

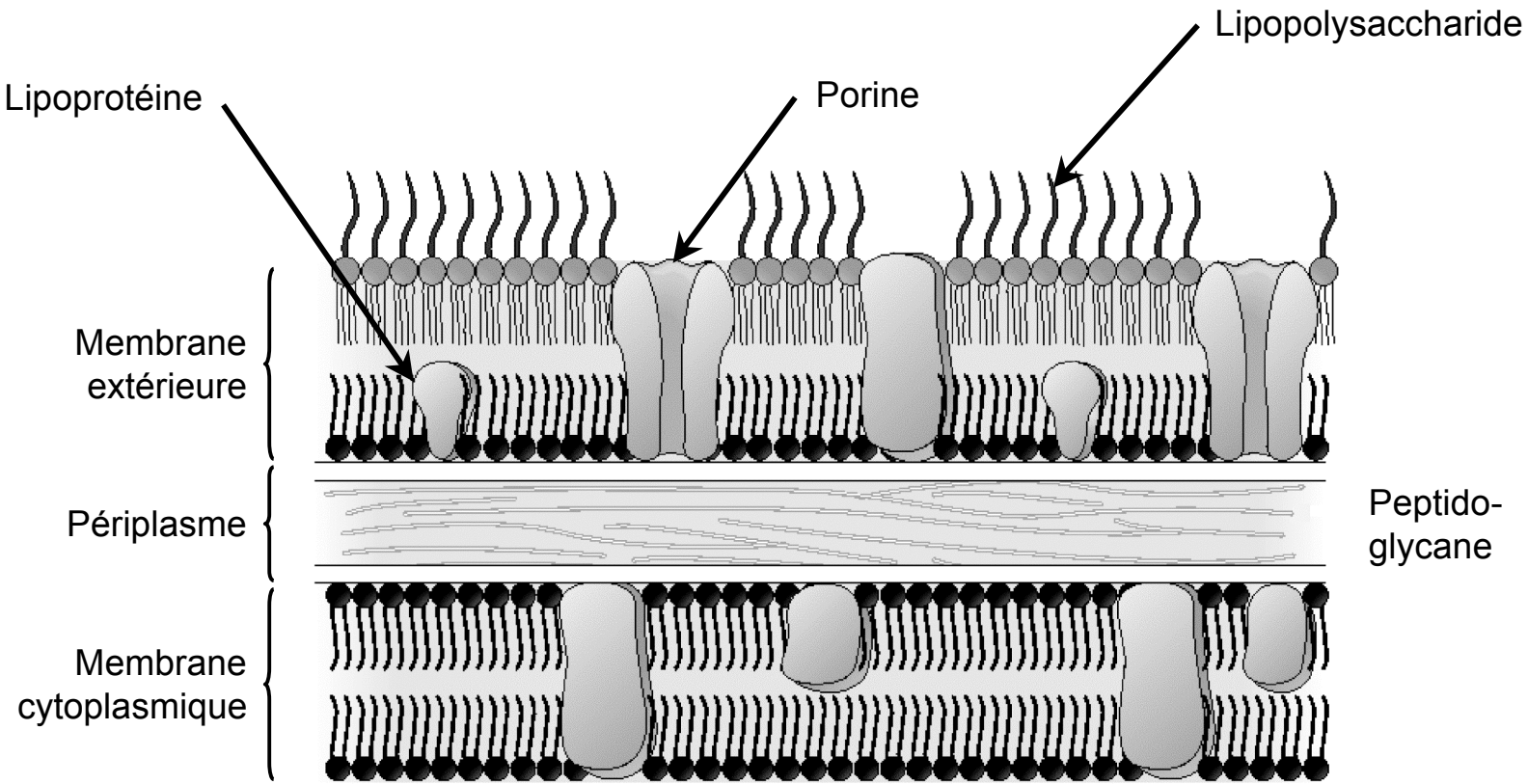
Arbre obtenu

Attraction de la longue branche de la séquence A par la longue branche de l'*outgroup* (O)

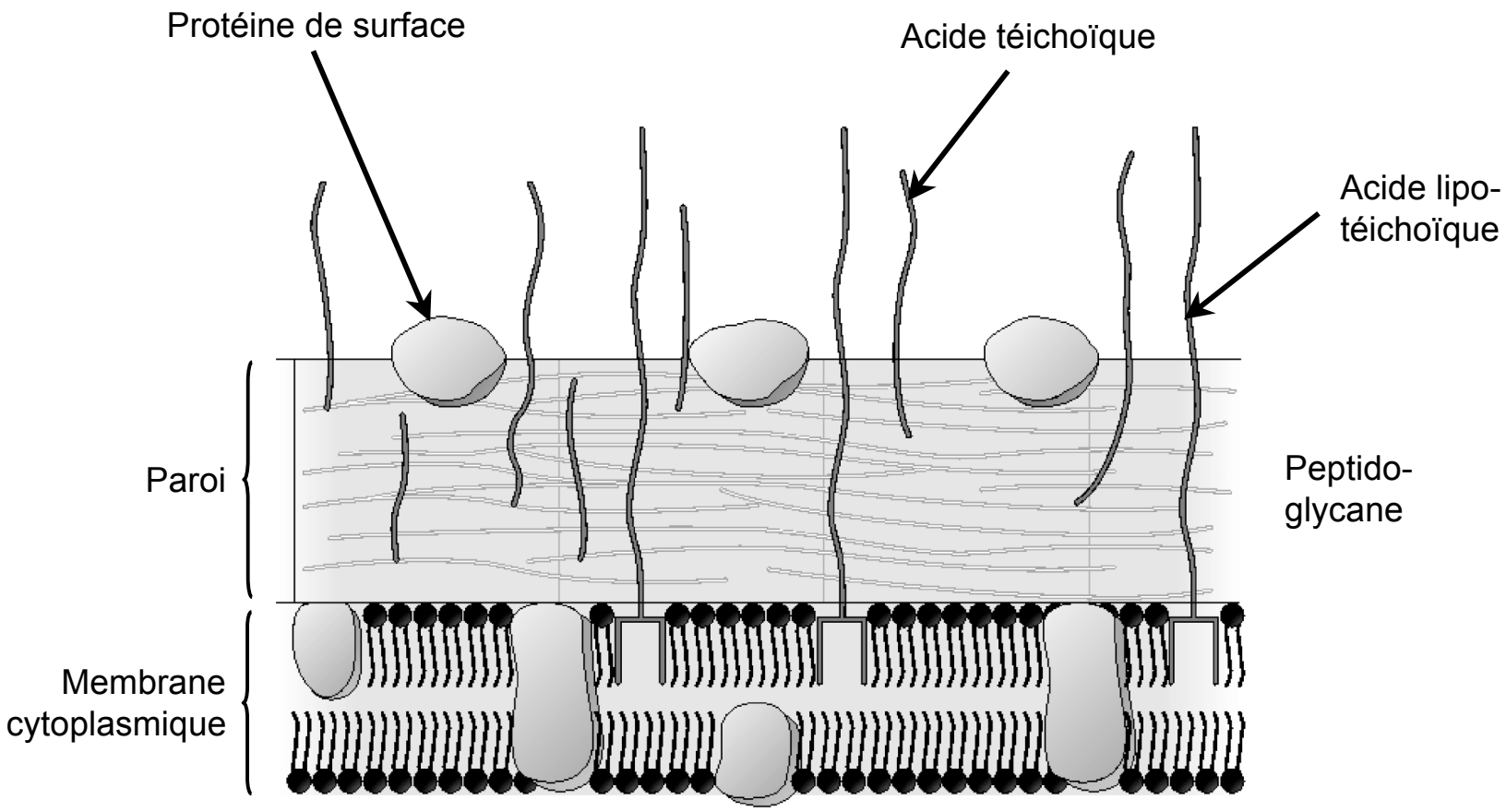
Principaux taxons bactériens

Division	Subdivision	Genres représentatifs
Protéobactéries	α-Protéobactéries β-Protéobactéries δ-Protéobactéries ε-Protéobactéries γ-Protéobactéries	*Agrobacterium, Rickettsia* *Neisseria*, *Ralstonia* *Myxobacterium* *Helicobacter*, *Campylobacter* *Escherichia*, *Buchnera*, *Pseudomonas*
Gram positives	Haut G+C Bas G+C	*Actinomyces*, *Streptomyces*, *Mycobacterium* *Bacillus*, *Clostridium*, *Mycoplasma*
Cyanobactéries et apparentées		*Nostoc*, *Synechocystis*
Spirochète et apparentées	Spirochètes Leptospiras	*Treponema*, *Borrelia* *Leptonema*, *Leptospira*
Bactéries vertes sulfureuses		*Chlorobium*, *Chloroherpeton*
Cytophagales, Flavobactéries et Bactéroïdes (CFB)	Bactéroïdes Flavobactéries	*Bacteroides*, *Fusobacterium* *Cytophaga, Flavobacterium*
Planctomycètes et apparentées	Groupe des Planctomycètes Groupe des Thermophiles	*Planctomyces*, *Pasteuria, Pirellula* *Isosphaera*
Chlamydiales		*Chlamydia*
Micrococcus radiorésistants et apparentées	Deinococcales Groupe des Thermophiles	*Deinococcus* *Thermus*
Bactéries vertes non sulfureuses	Groupe des Chloroflexus Groupe des Thermomicrobium	*Chloroflexus*, *Herpetosiphon* *Thermomicrobium*
Aquificales et Thermotogales	Aquificales Thermotogales	*Aquifex*, *Hydrogenobacter* *Thermotoga*, *Geotoga*, *Thermopallium*

Paroi des Gram négatives



Paroi des Gram positives

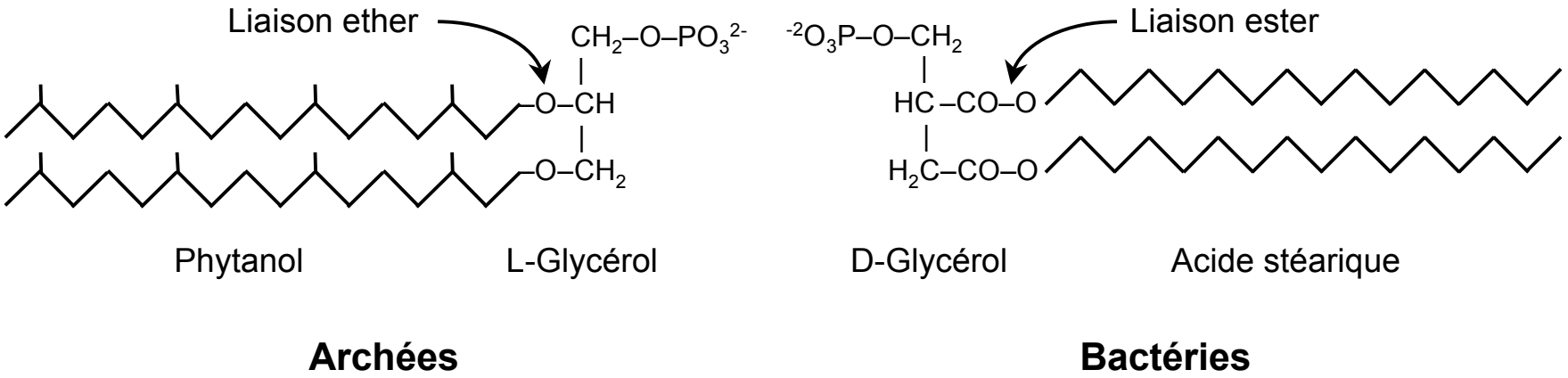


Principaux taxons archéens

Division	Subdivision	Genres représentatifs
Crénarchées	Thermoprotéales	*Thermoproteus, Pyrobaculum, Thermofilum*
	Sulfolobales	*Sulfolobus*, *Acidianus*
	Désulfurococcales	*Aeropyrum, Desulfurococcus*
	Cénarchéales	*Cenarchaeum*
	Caldisphérales	*Caldisphaera*
Euryarchées	Méthanobacteria	*Methanobacterium, Methanothermobacter*
	Méthanococci	*Methanococcus, Methanothermococcus*
	Halobactéria	*Halobacterium, Halococcus*
	Thermoplasmata	*Thermoplasma, Ferroplasma*
	Thermococci	*Pyrococcus, Thermococcus*
	Archaeoglobi	*Archaeoglobus*
	Methanopyri	*Methanopyrus*
	Methanomicrobia	*Methanosarcina, Methanoculleus*
Korarchées		Séquences environementales
Nanoarchées [?]		*Nanoarchaeum*

Particularités des archées

- Structure particulière de la paroi :
  - Pas de peptidoglycane.
  - Diglycérides membranaires spécifiques :
    - L-Glycérol au lieu de D-Glycérol.
    - Isoprènes au lieu d'acides gras.
    - Liaisons avec le glycérol de type ether au lieu d'ester.



Choix d'un marqueur

- Gènes évoluant très lentement (temps de divergence entre 10^7 et 10^9 ans) :
  - Éviter la saturation et les artefacts d'attraction de longues branches.
- Pas de paralogues (ou paralogues facilement identifiables).
- Pas de transferts horizontaux.
- Présents dans l'ensemble des organismes étudiés.

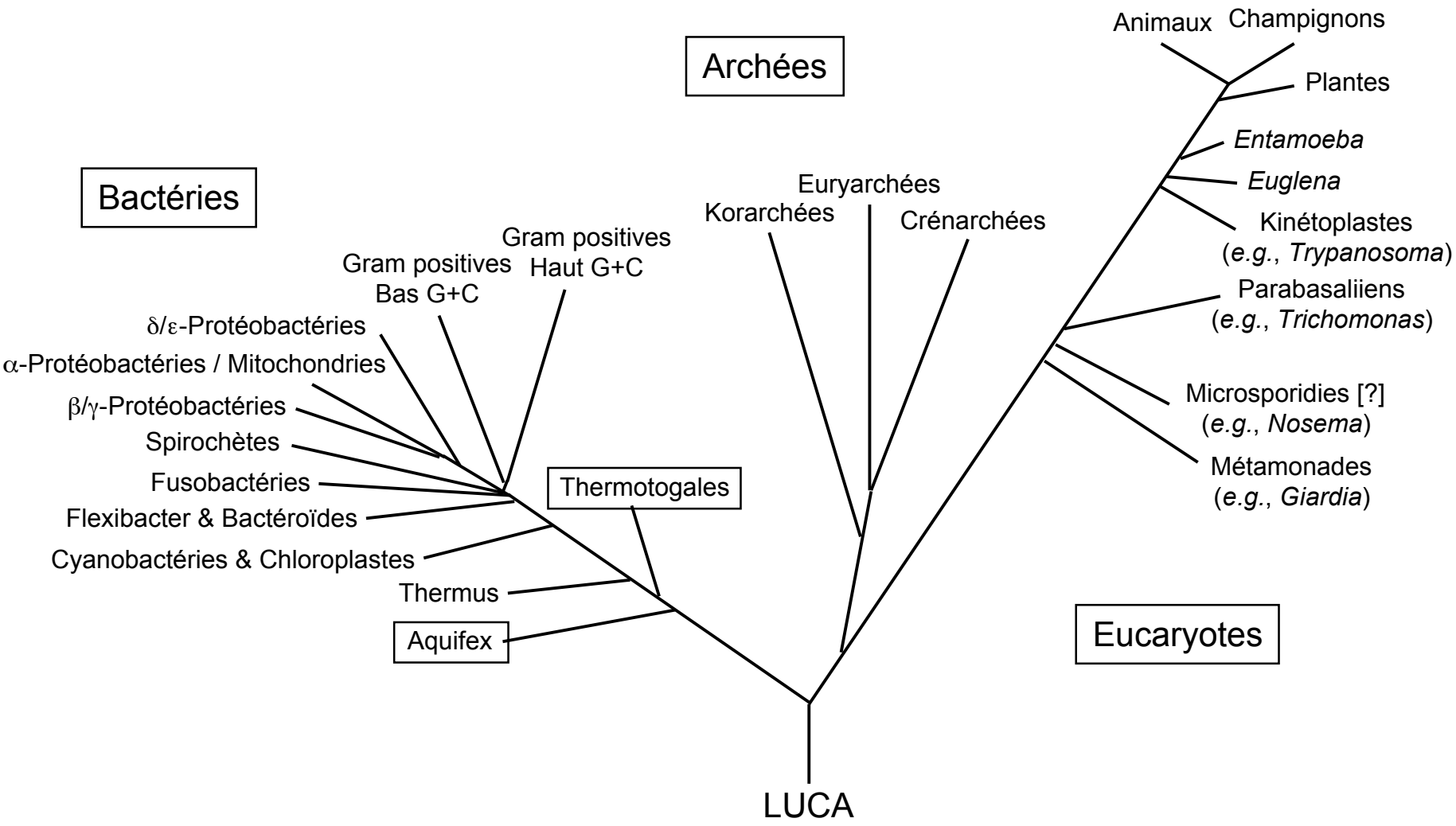
Marqueurs possibles

- Seul un petit nombre de gènes répondent aux critères précédents :
  - ARNr de la petite ou de la grande sous-unité du ribosome.
  - Protéines *heat-shock* (*e.g.*, Hsp70).
  - Facteurs d'élongation de la traduction.
  - Protéines ribosomiques.
- Résultats fréquemment incongruents entre eux.

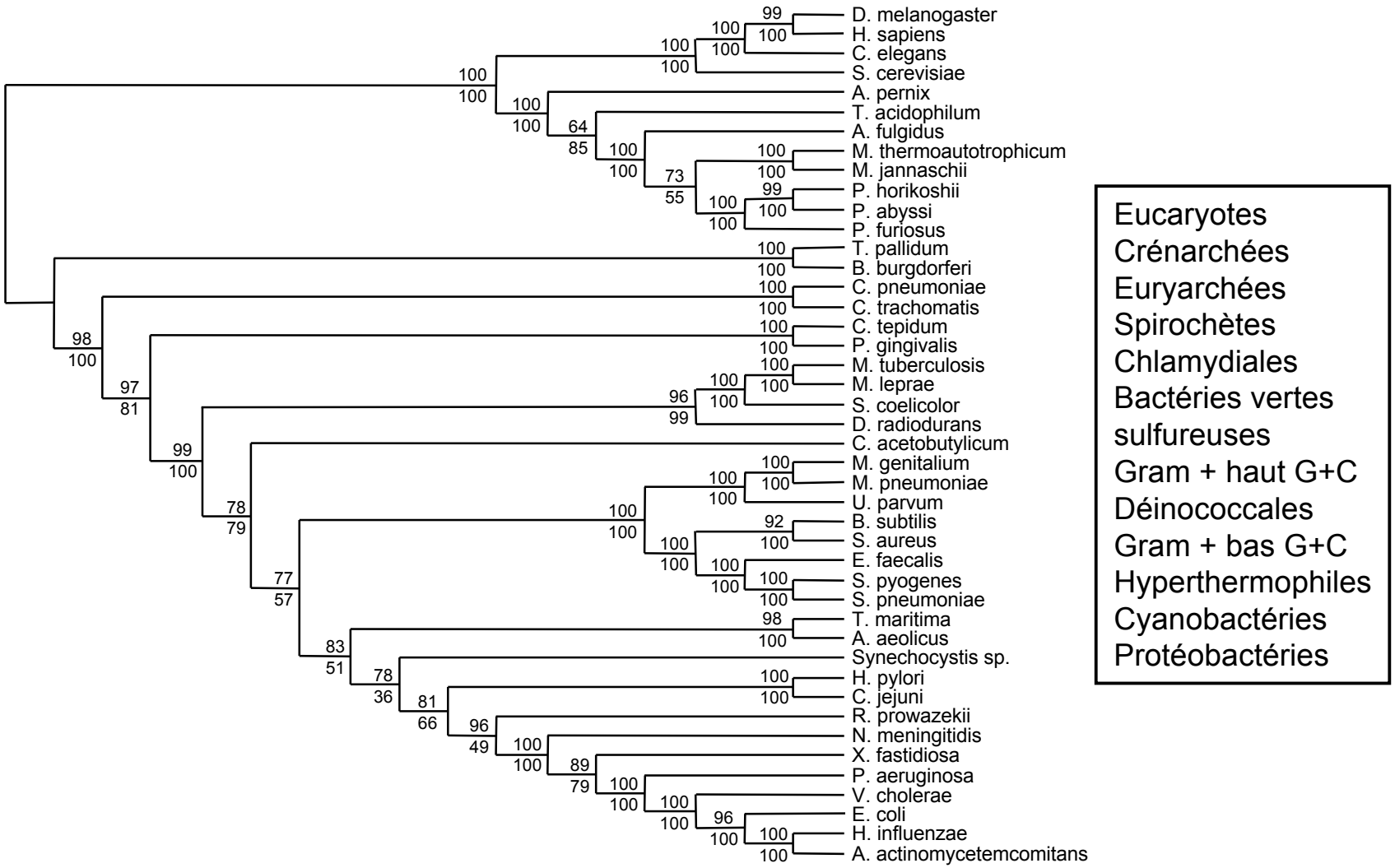
Phylogénomique

- Utilisation de génomes complets plutôt que des gènes individuels.
- Minimisation de l'effet des paralogies cachées, des transferts et des erreurs.
- Intégration d'une plus grande quantité de signal.
- Différentes approches ont été tentées :
  - Codage en présence/absence des gènes.
  - Concaténation d'orthologues présents dans de nombreuses espèces.
  - Super-arbres intégrant l'information provenant d'un ensemble d'arbres individuels.

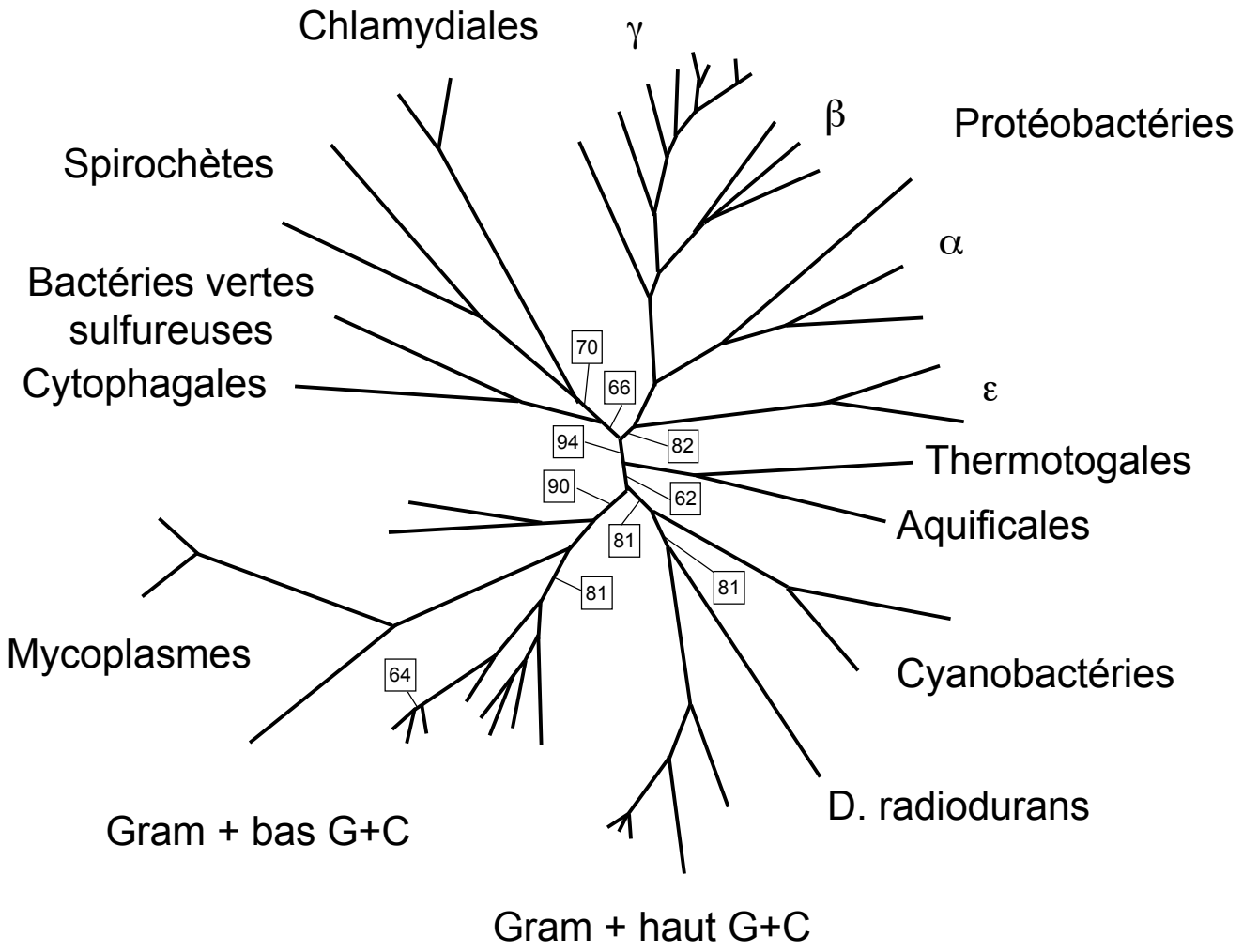
L’arbre de Brown (1997)



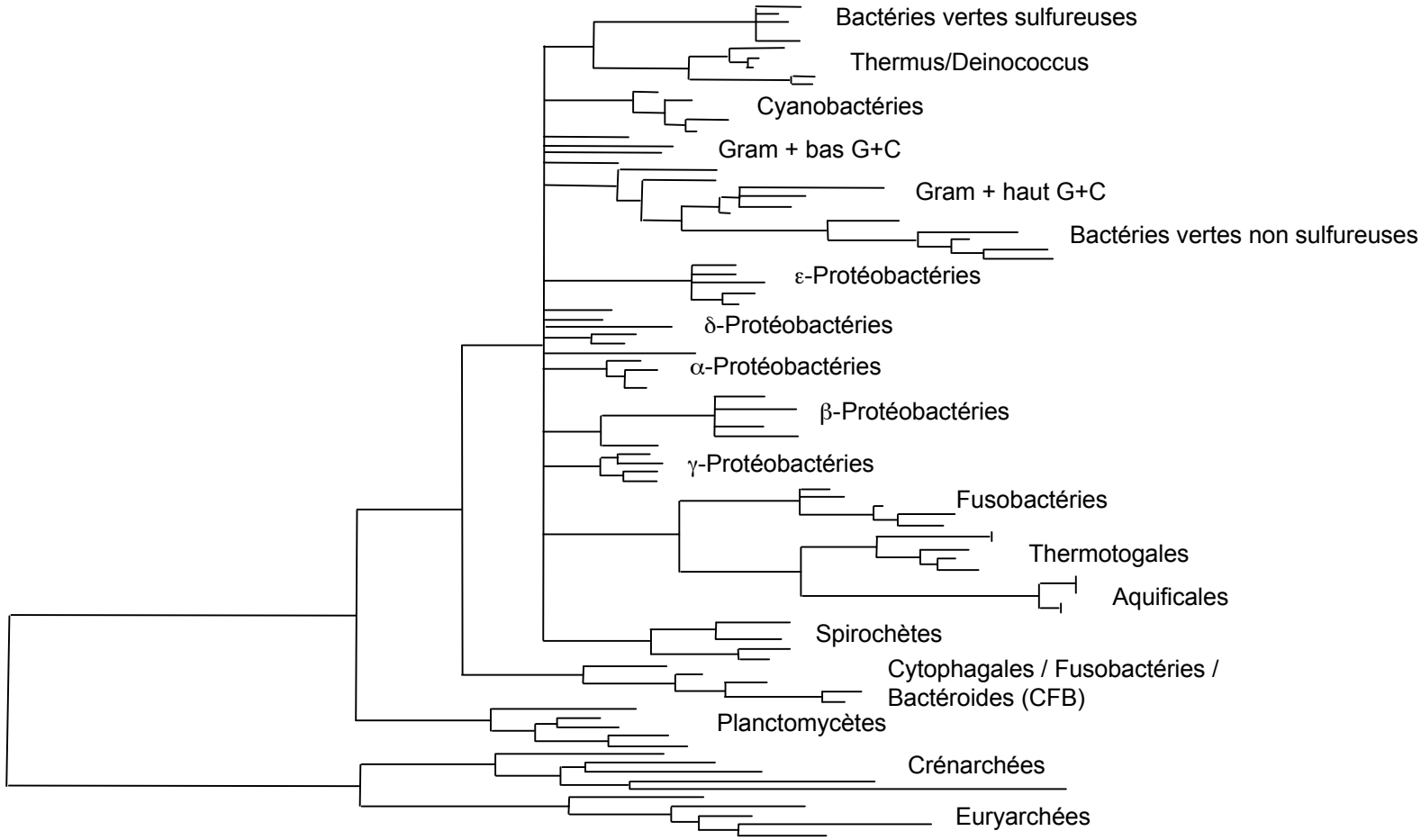
L’arbre de Brown (2001)



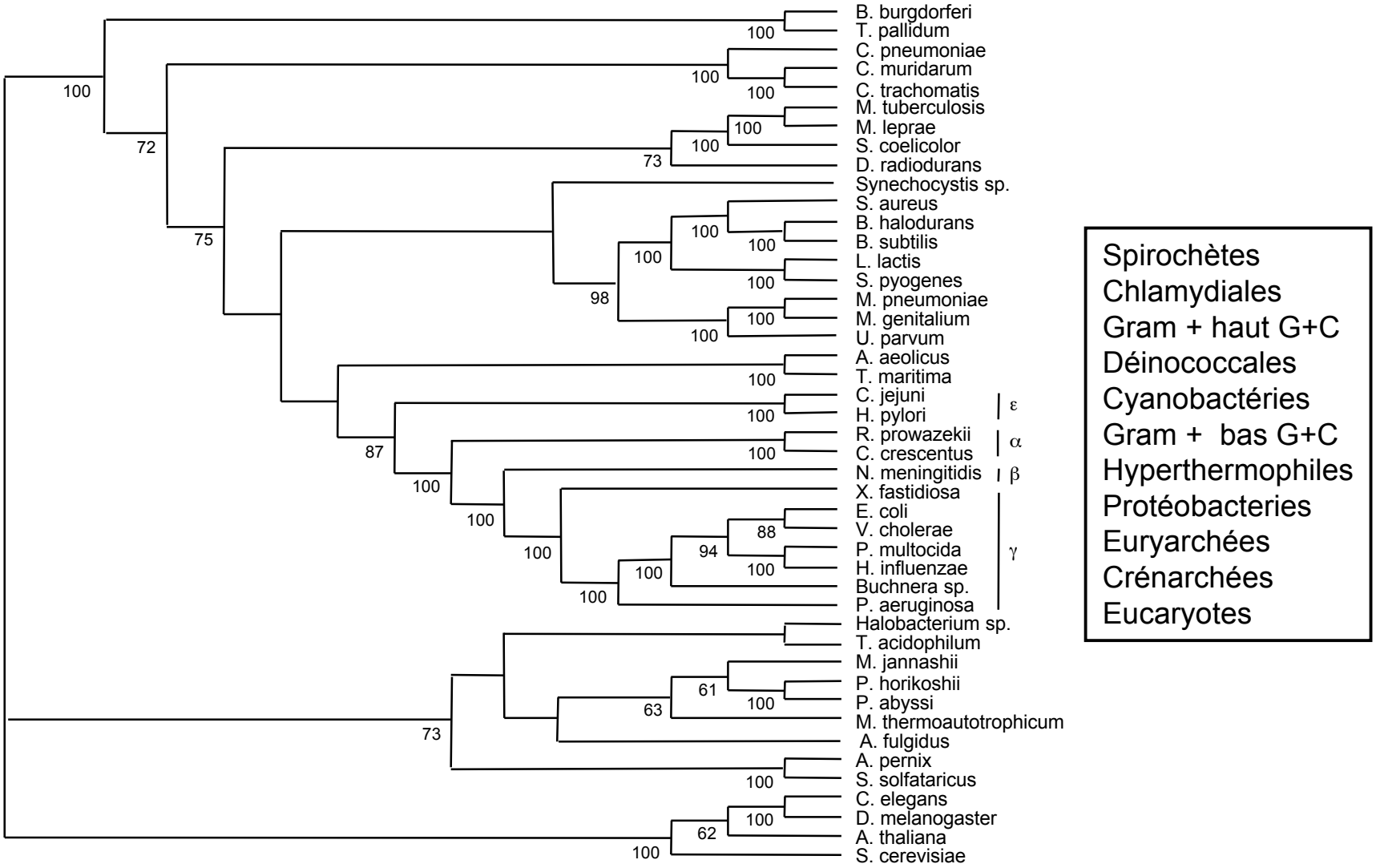
L’arbre de Brochier (2001)

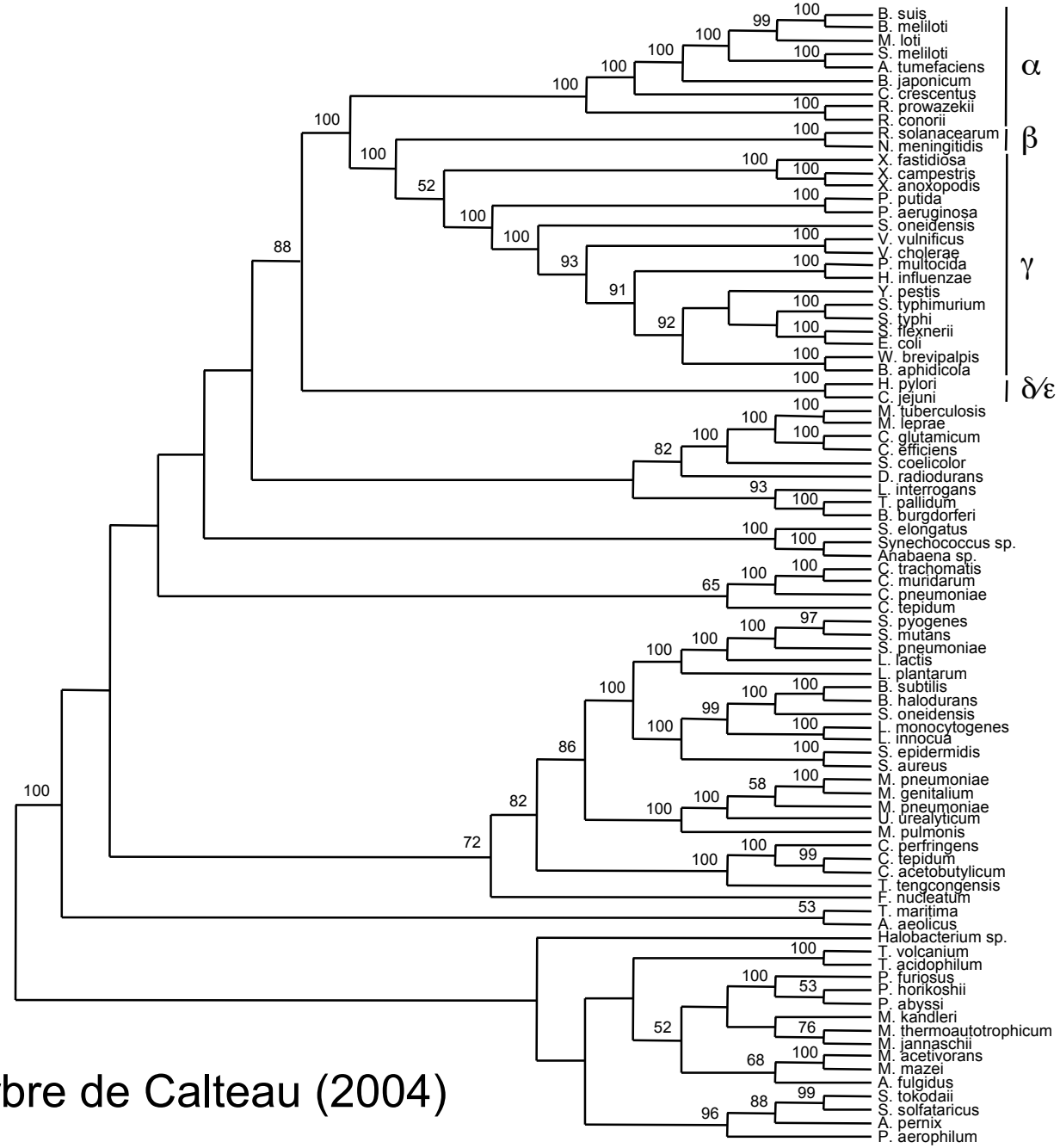


L’arbre de Brochier (2002)



L’arbre de Daubin (2002)





L'arbre de Calteau (2004)

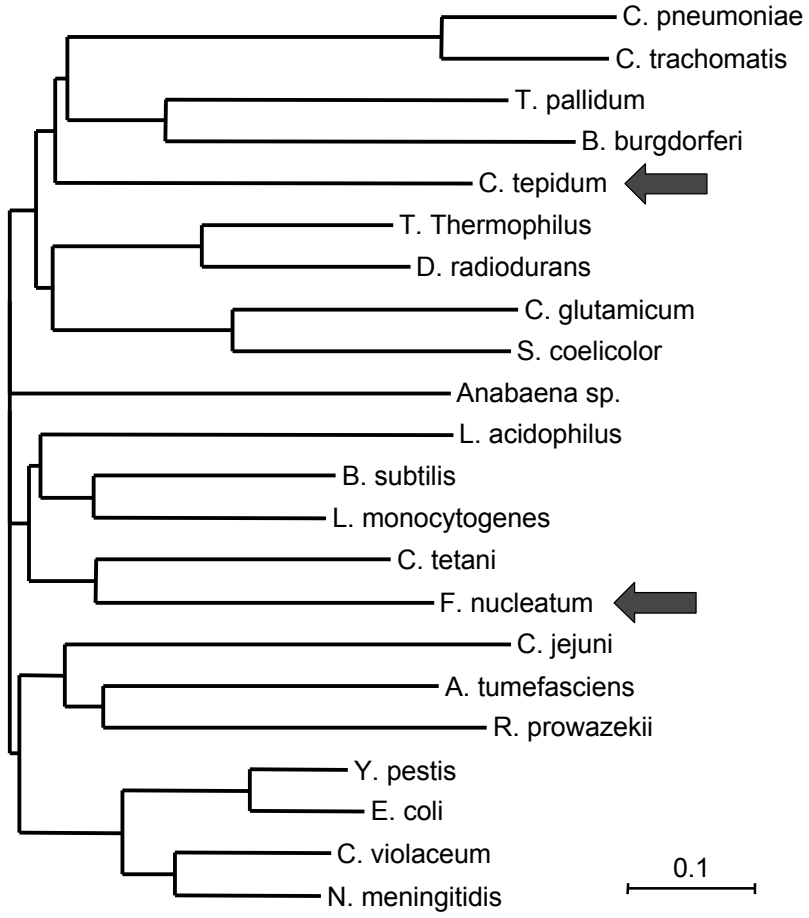
Protéobacteries
Gram + haut G+C
Déinococcales
Spirochètes
Cyanobactéries
Chlamydiales
Chlorobiales
Gram + bas G+C
Fusobactéries (CFB)
Hyperthermophiles
Euryarchées
Crénarchées

Que retenir ?

- Monophylie de chacun des trois grands domaines du vivant.
- Monophylie des grands groupes bactériens et archéens définis par Woese.
- Monophylie des différentes subdivisions des Protéobactéries.
- Polyphylie des bactéries Gram + apparemment établie.
- Positionnement variable de la plupart des grandes divisions taxonomiques les unes par rapport aux autres.

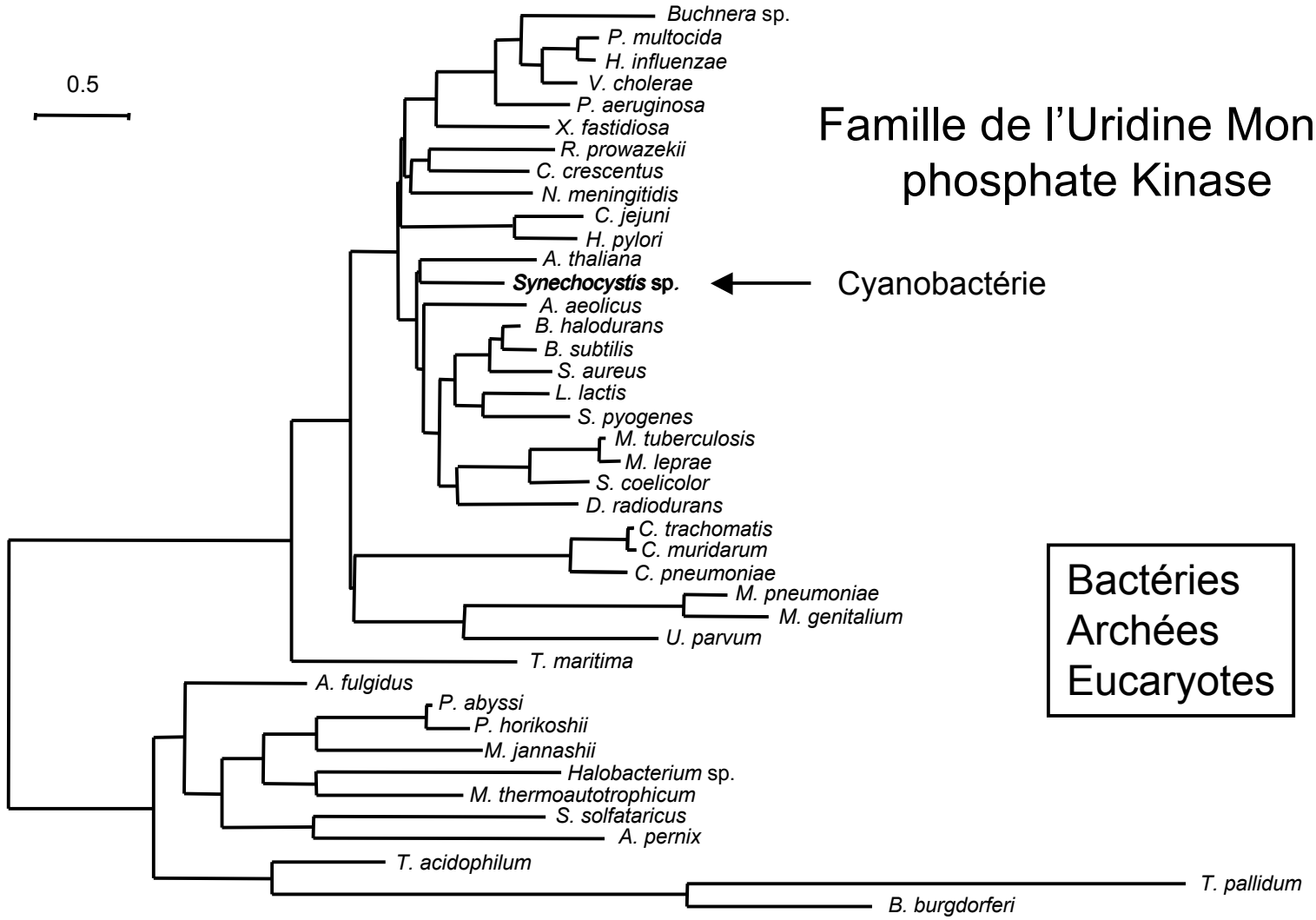
Importance de la référence

- Transferts inférés en fonction de la référence.
- Cas du gène *ruvB* :
  - Arbre de Brown (2001) :
    - Transfert depuis l'ancêtre commun spirochètes / chlamydiales → *C. tepidum*.
  - Arbre de Brochier (2001) :
    - Transfert depuis Gram + bas G+C → *F. nucleatum*.
  - Arbre de Calteau (2004) :
    - Pas de transfert.

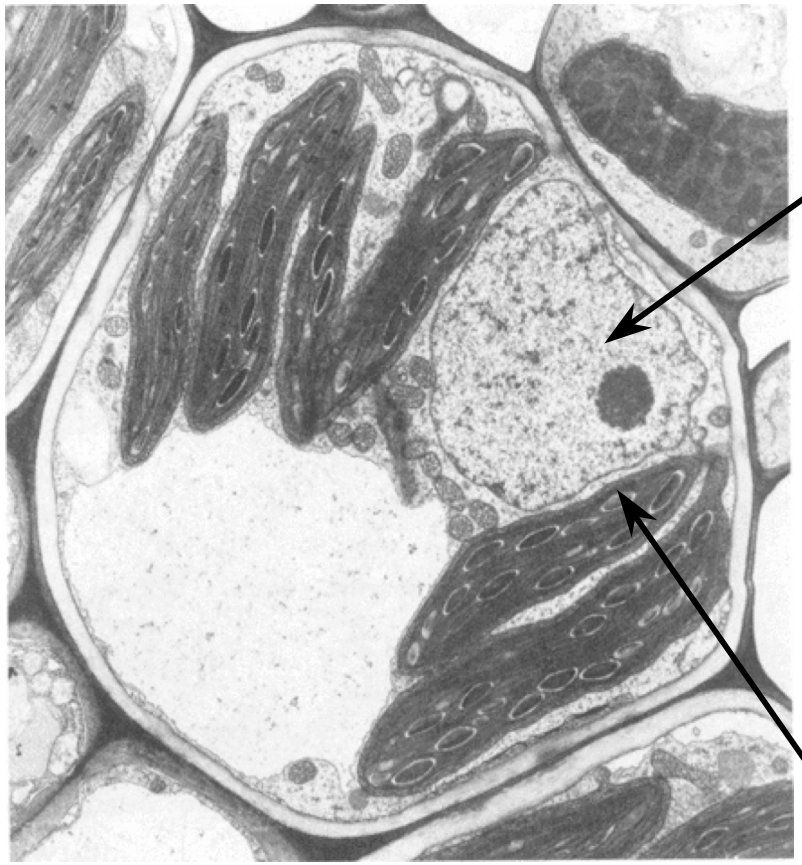


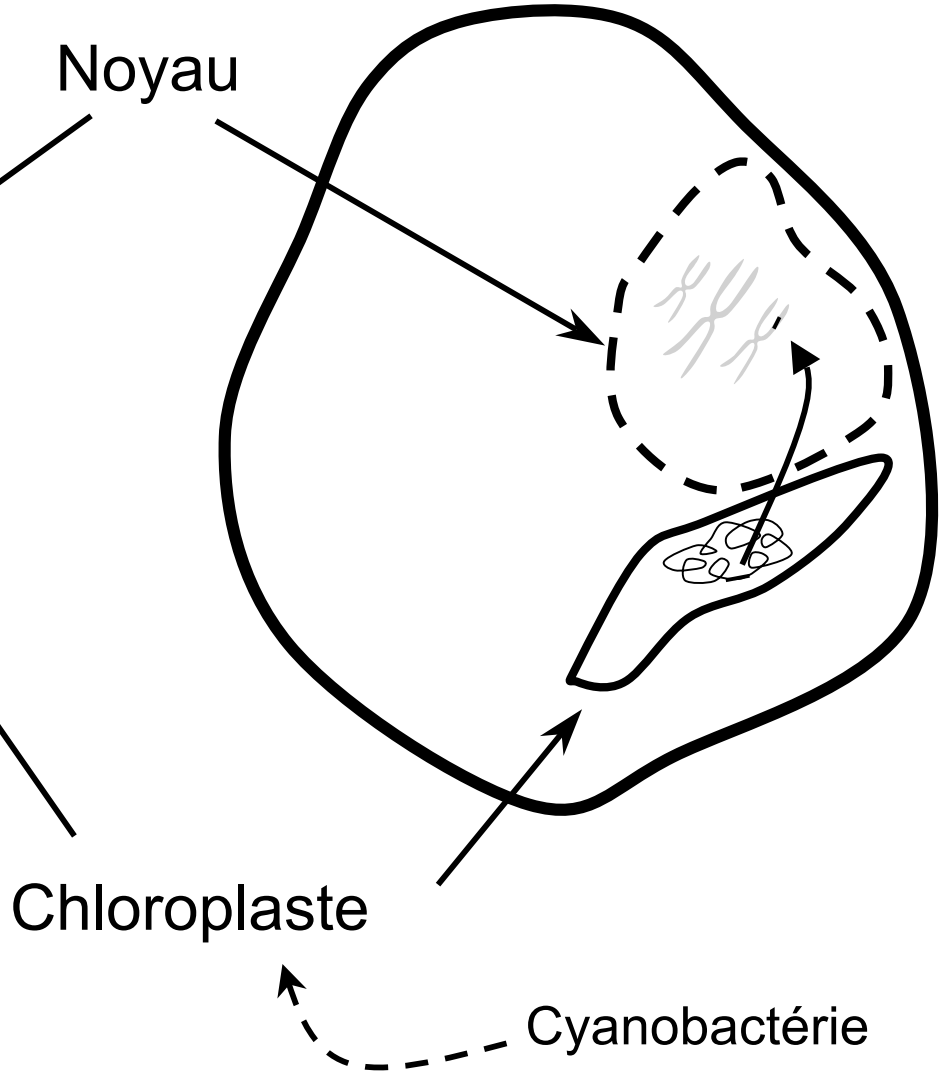
DNA hélicase RuvB

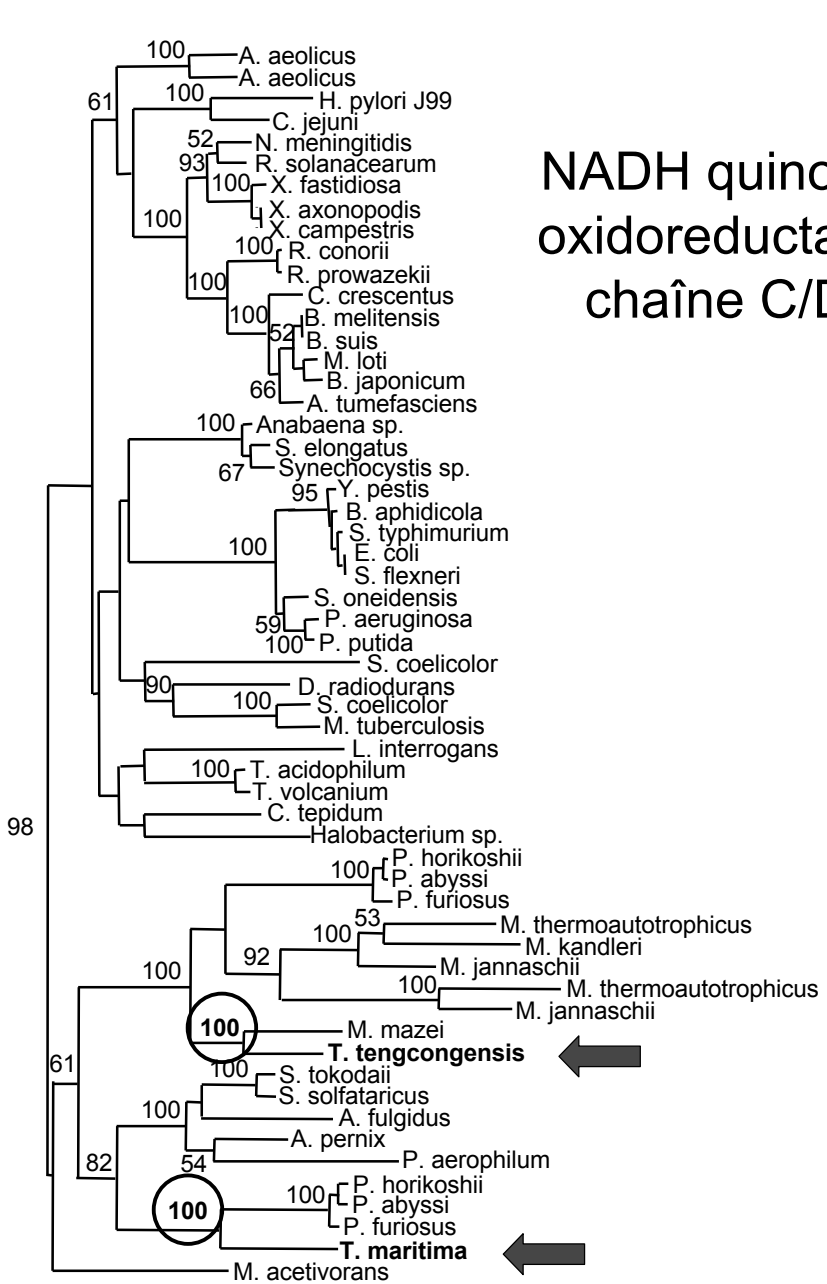
Cas du gène *pyrH*



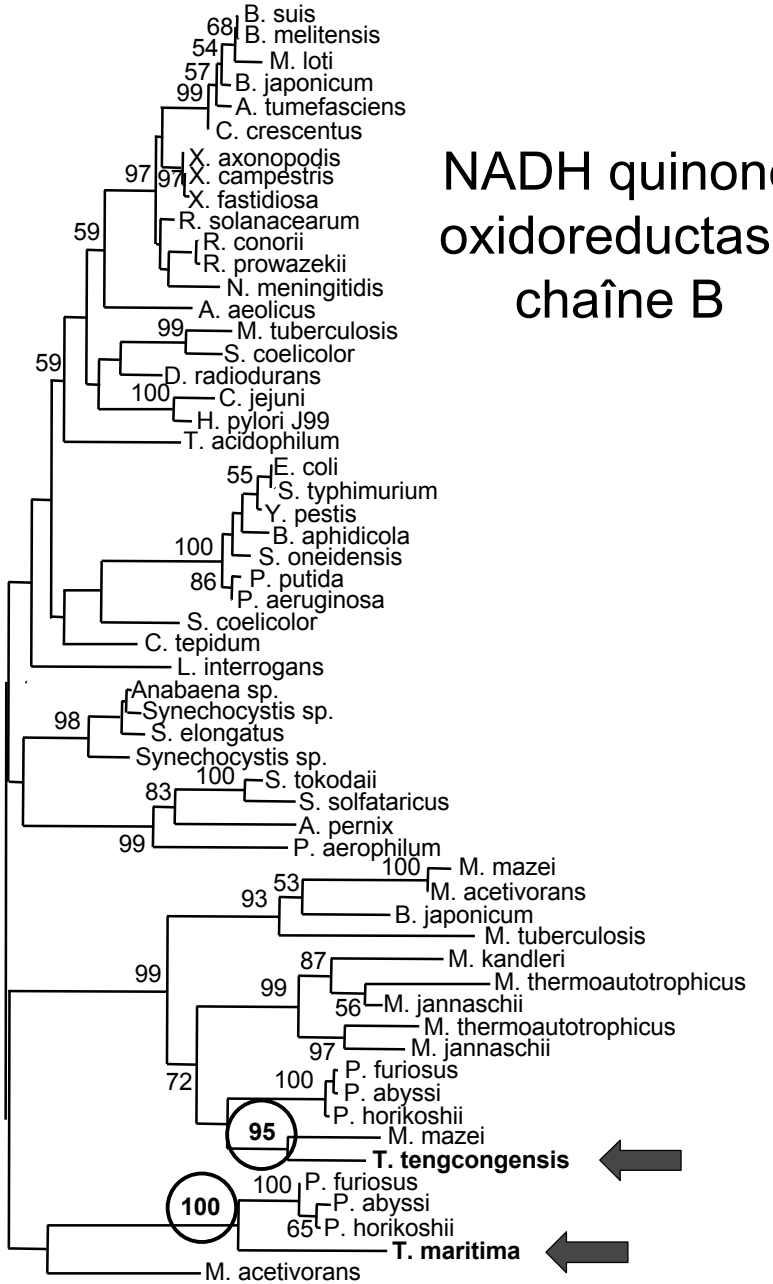
Endosymbiose et transferts







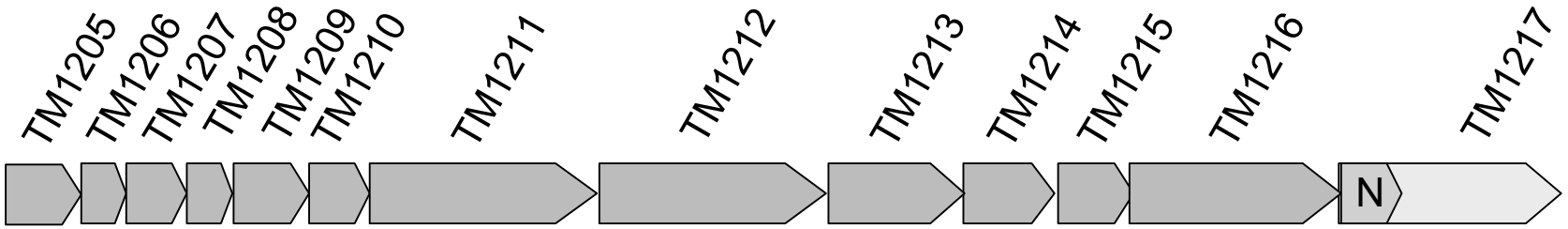
NADH quinone oxidoreductase chaîne C/D



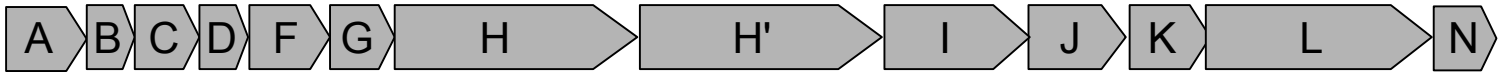
NADH quinone oxidoreductase chaîne B

L'opéron *mbx*

Thermotoga maritima



Thermococcales

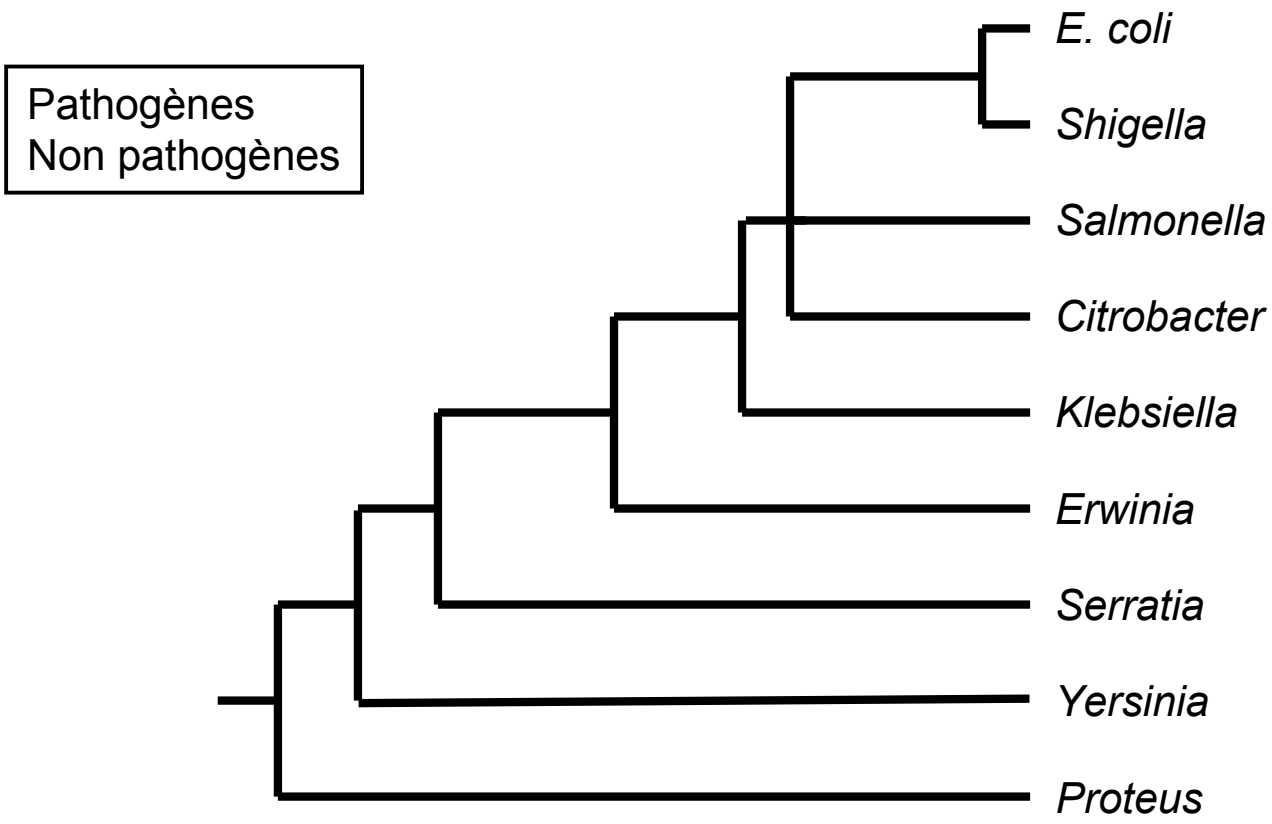


- Transfert d'un opéron entier de 13 gènes :
  - Insertion à l'intérieur de la séquence d'une glutamate synthase.
- Opéron fonctionnel chez *T. maritima* ?
  - Quel rôle dans cet organisme ?

Transferts et pathogénicité

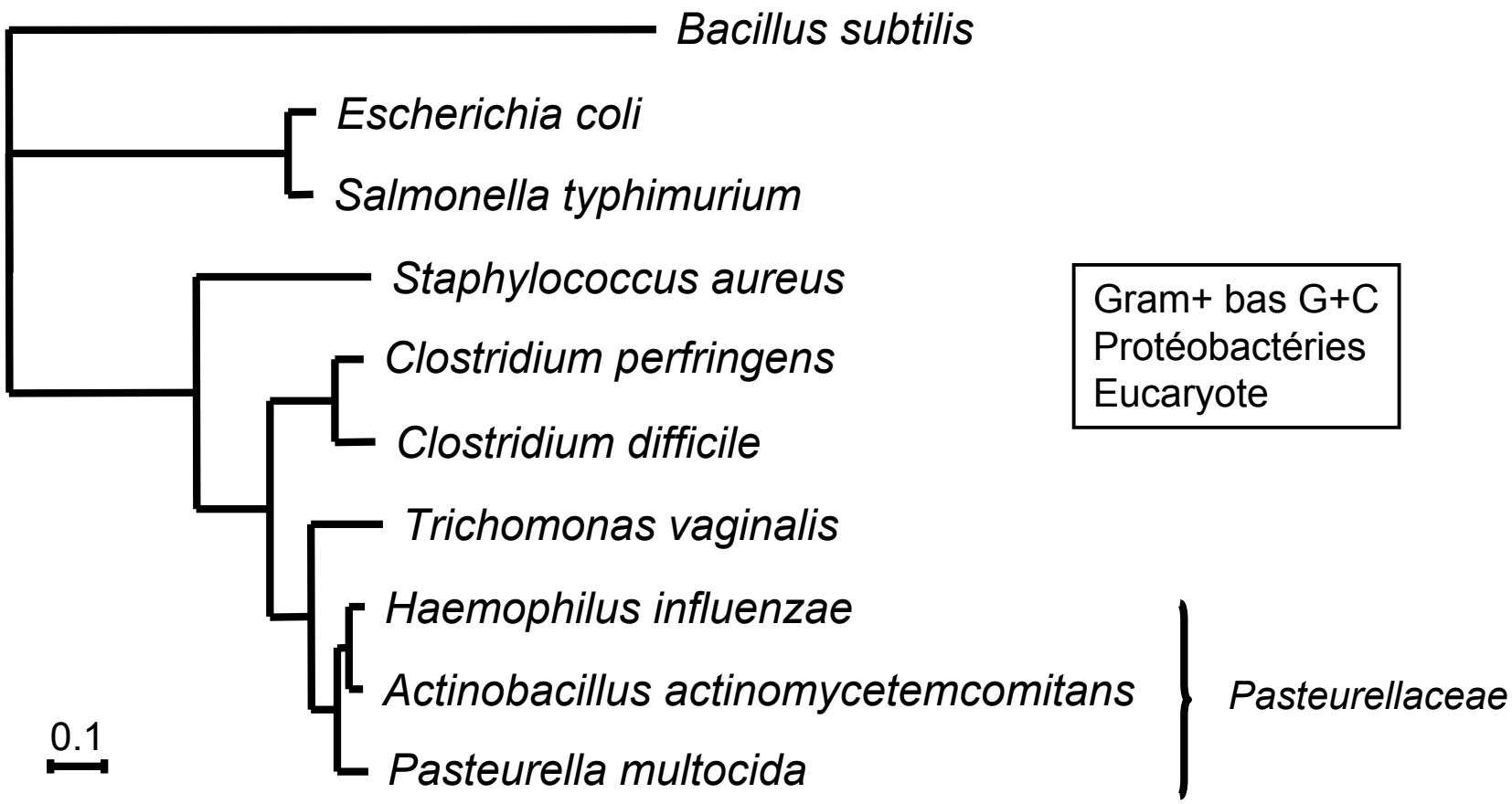
- Transposons :
  - Gènes de l'entérotoxine ST d'*E. coli*.
- Prophages :
  - Toxines de l'*E. coli* entéro-hémorragique.
  - Gènes de la toxine diphtérique ou du choléra.
  - Toxines botuliniques.
- Plasmides :
  - *Shigella*, *Salmonella*, *Yersinia*.
- Ilôts de pathogénicité :
  - *E. coli* uropathogène, *S. typhimurium*, *Yersinia*, *H. pylori*, *V. cholerae*.

Le cas des entérobactéries



La virulence est un état *dérivé*, obtenu par l'acquisition de gènes déterminant cet état

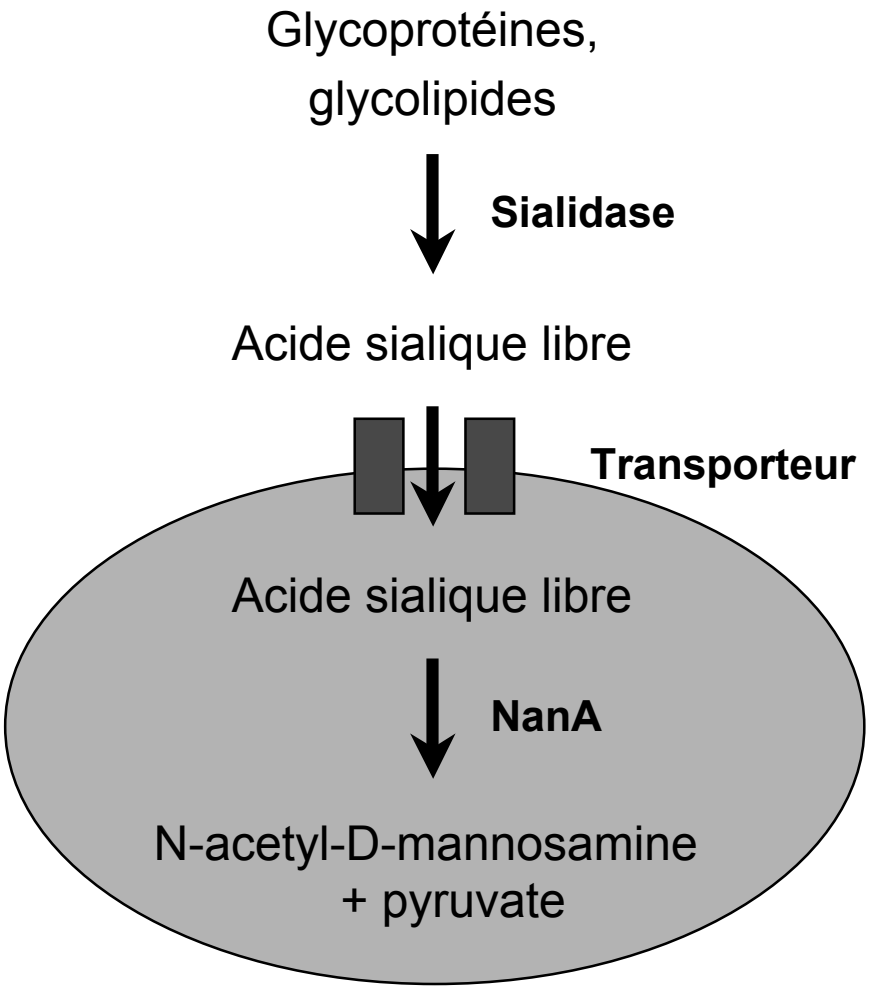
Transfert bactérie-eucaryote



Phylogénie de la *N*-acetylneuraminate lyase (NanA)

NanA et pathogénicité ?

- Enzyme impliquée dans le métabolisme de l'acide sialique.
- Rôle chez les bactéries :
  - Parasitage des muqueuses des animaux à des fins nutritionnelles.
- Rôle inconnu chez *T. vaginalis*.



Méthodes intrinsèques

- Utilisent exclusivement l'information stockée au sein du génome étudié.
- Se basent principalement sur la composition en codons ou en nucléotides des gènes :
  - Utilisent des indices (univariés ou multivariés) d'usage du code.
  - Un gène transféré présentera un usage du code différent de celui des gènes « natifs ».
  - Nécessité de prendre en compte les facteurs intervenant dans la plasticité des génomes.

Facteurs connus

- Les facteurs influençant la composition en codons chez les bactéries sont multiples :
  - Contenu global en G+C du génome.
  - Sélection traductionnelle.
  - Localisation sur l'un ou l'autre des deux brins du chromosome.
  - Distance à l'origine de réplication.
  - Composition en acides aminés.

Le code génétique standard

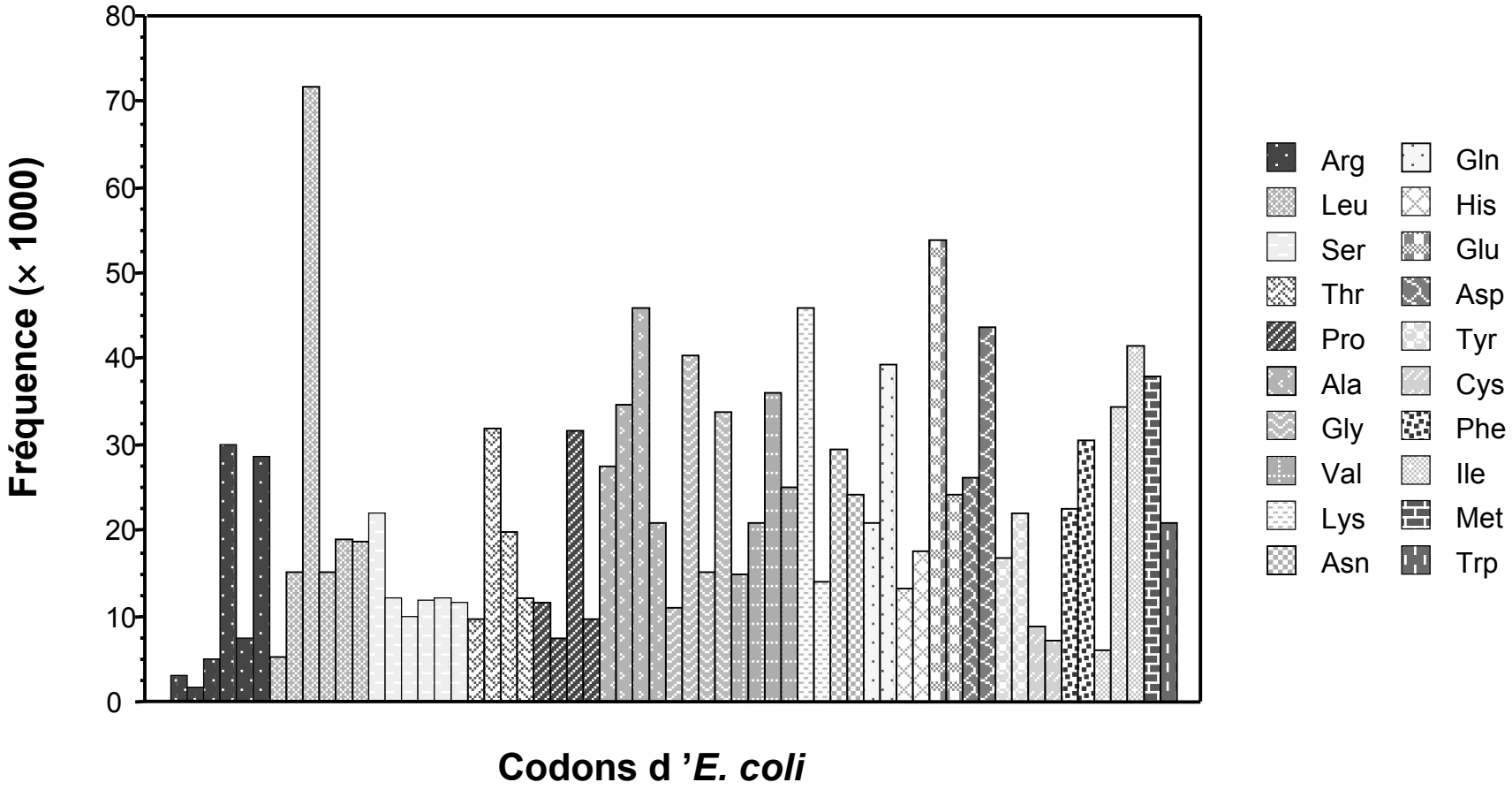
I \ II	U	C	A	G	III
U	UUU Phe F	UCU Ser S	UAU Tyr Y	UGU Cys C	**U**
	UUC Phe F	UCC Ser S	UAC Tyr Y	UGC Cys C	**C**
	UUA Leu L	UCA Ser S	UAA Stop	UGA Stop	**A**
	UUG Leu L	UCG Ser S	UAG Stop	UGG Trp W	**G**
C	CUU Leu L	CCU Pro P	CAU His H	CGU Arg R	**U**
	CUC Leu L	CCC Pro P	CAC His H	CGC Arg R	**C**
	CUA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R	**A**
	CUG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R	**G**
A	AUU Ile I	ACU Thr T	AAU Asn N	AGU Ser S	**U**
	AUC Ile I	ACC Thr T	AAC Asn N	AGC Ser S	**C**
	AUA Ile I	ACA Thr T	AAA Lys K	AGA Arg R	**A**
	AUG Met M	ACG Thr T	AAG Lys K	AGG Arg R	**G**
G	GUU Val V	GCU Ala A	GAU Asp D	GGU Gly G	**U**
	GUC Val V	GCC Ala A	GAC Asp D	GGC Gly G	**C**
	GUA Val V	GCA Ala A	GAA Glu E	GGA Gly G	**A**
	GUG Val V	GCG Ala A	GAG Glu E	GGG Gly G	**G**

Biais au niveau d'un gène

UUU Phe	6	UCU Ser	5	UAU Tyr	4	UGU Cys	0
UUC Phe	10	UCC Ser	6	UAC Tyr	12	UGC Cys	3
UUA Leu	8	UCA Ser	8	UAA Ter	*	UGA Ter	*
UUG Leu	6	UCG Ser	10	UAG Ter	*	UGG Trp	12
CUU Leu	6	CCU Pro	5	CAU His	2	CGU Arg	7
CUC Leu	9	CCC Pro	5	CAC His	3	CGC Arg	6
CUA Leu	5	CCA Pro	4	CAA Gln	9	CGA Arg	6
CUG Leu	2	CCG Pro	3	CAG Gln	9	CGG Arg	3
AUU Ile	1	ACU Thr	11	AAU Asn	2	AGU Ser	4
AUC Ile	8	ACC Thr	5	AAC Asn	15	AGC Ser	3
AUA Ile	7	ACA Thr	5	AAA Lys	5	AGA Arg	3
AUG Met	7	ACG Thr	6	AAG Lys	9	AGG Arg	4
GUU Val	8	GCU Ala	6	GAU Asp	8	GGU Gly	15
GUC Val	7	GCC Ala	12	GAC Asp	5	GGC Gly	6
GUA Val	7	GCA Ala	7	GAA Glu	5	GGA Gly	2
GUG Val	9	GCG Ala	10	GAG Glu	12	GGG Gly	5

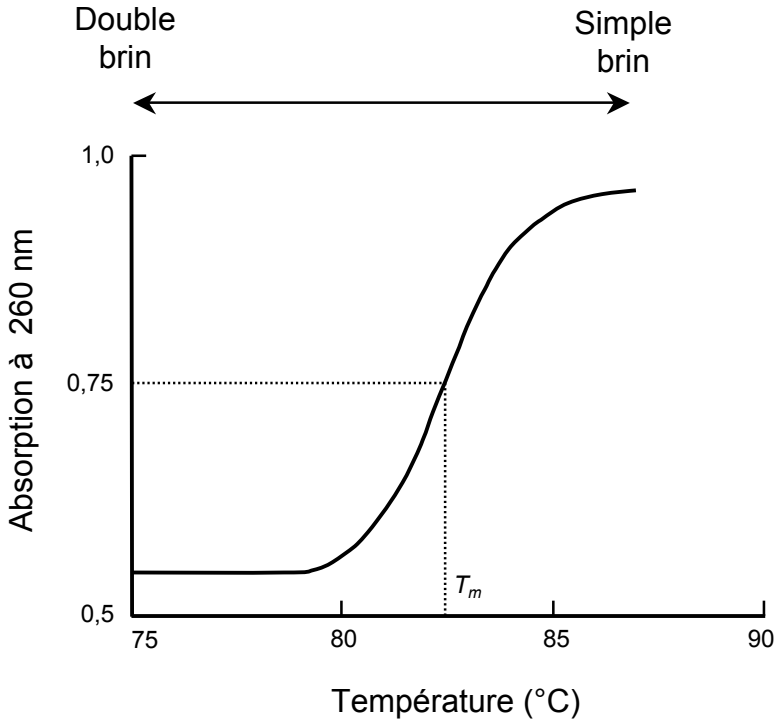
Protéine A du phage MS2 (Fiers *et al.*, 1975)

Biais au niveau d ’un génome



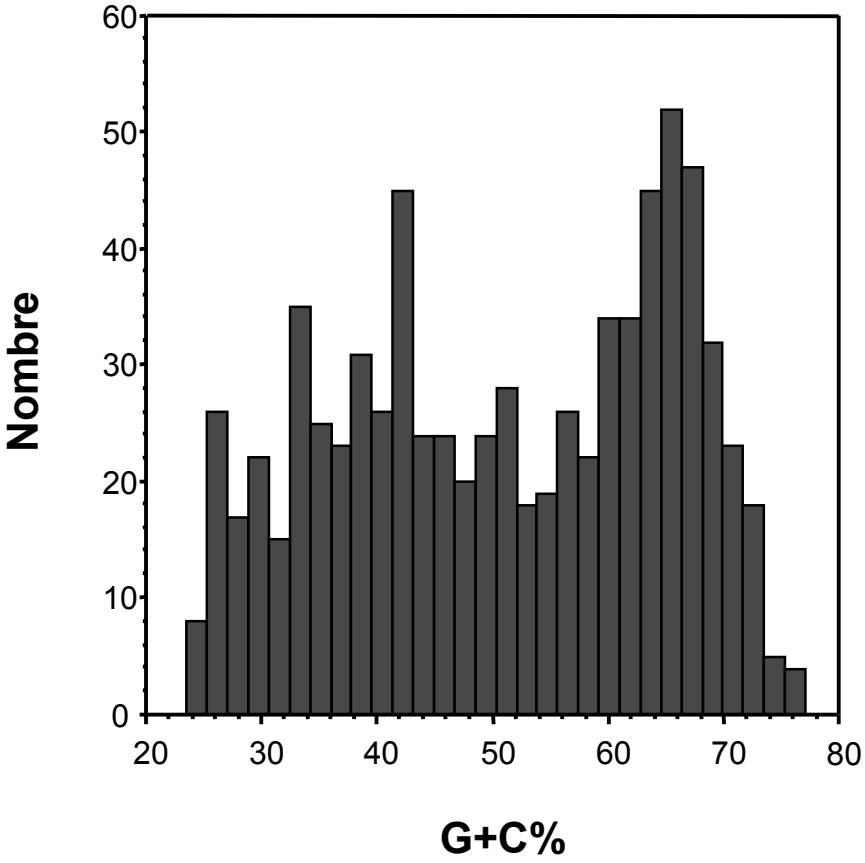
Contenu global en G+C

- Calculé en pourcentage de bases G+C :
  - Constitue une des premières mesures moléculaires appliquée à la systématique.
- Méthodes de dénaturation par la chaleur :
  - Mesure de la variation de l'absorption UV en fonction de la température.
  - La valeur du T_m est linéairement proportionnelle au G+C% de l'organisme.



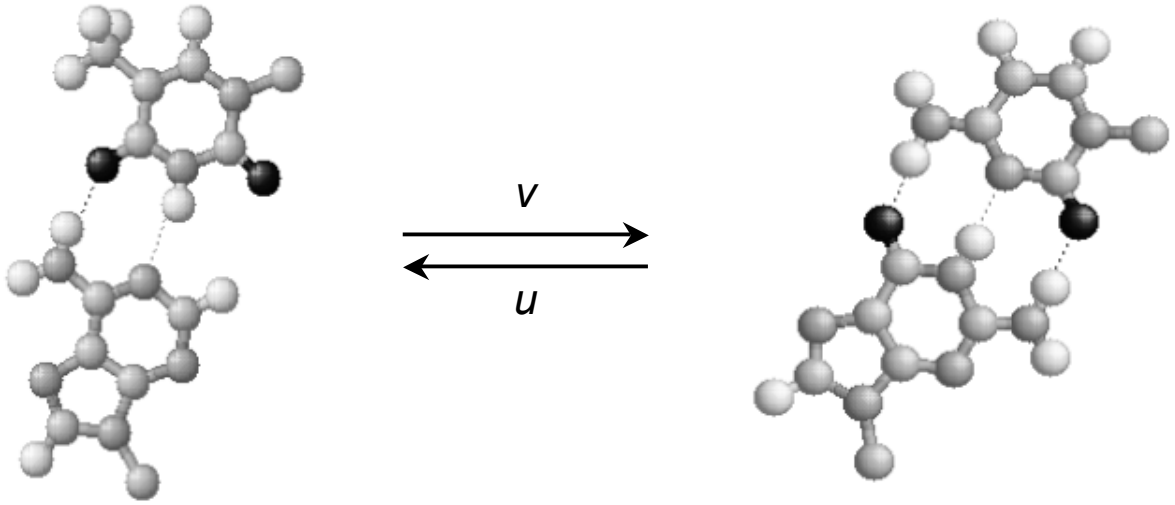
Distribution

- Mesure faite sur 772 bactéries.
- Minimum : *Mycoplasma capricolum* (24 %).
- Maximum : *Micrococcus luteus* (77 %).
- Variations de 5 % et 10 % au sein d'une espèce et d'un genre.



Source de la variation

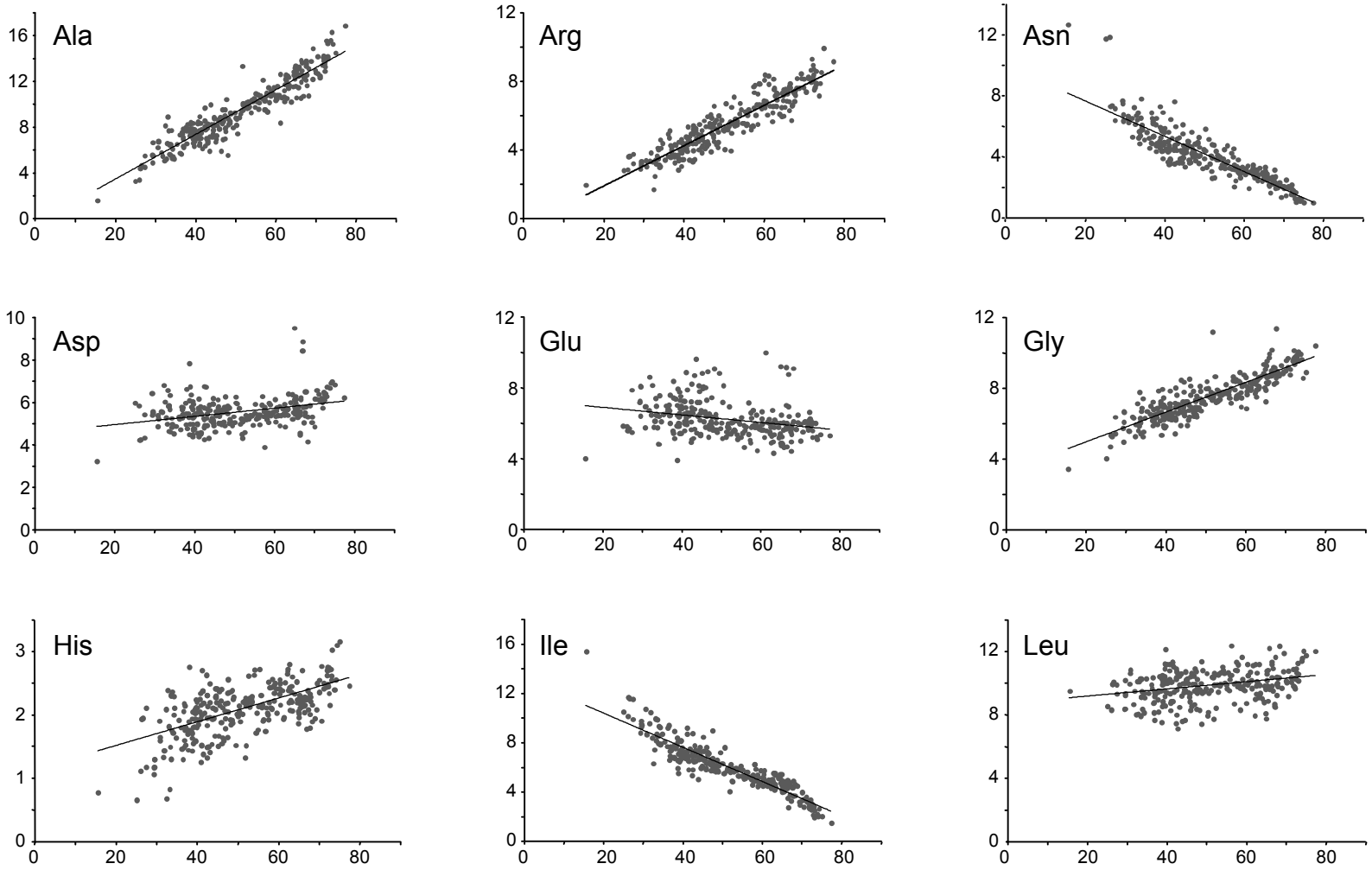
- Différences de pression de mutation entre les paires A•T et G•C.
- Le ratio *u*/*v* est différent d'une espèce à l'autre :
  - Distribution observée chez les bactéries.



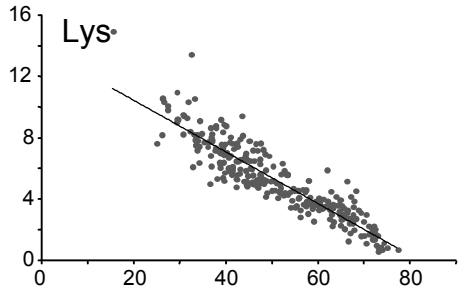
Paire A•T

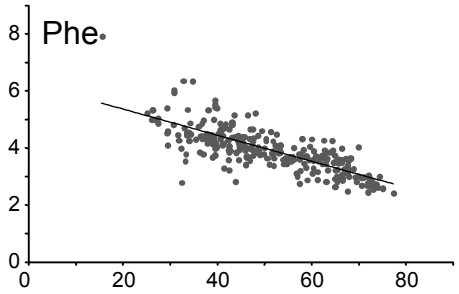
Paire G•C

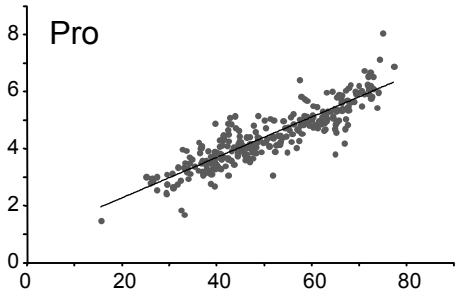
G+C global et contenu en AA

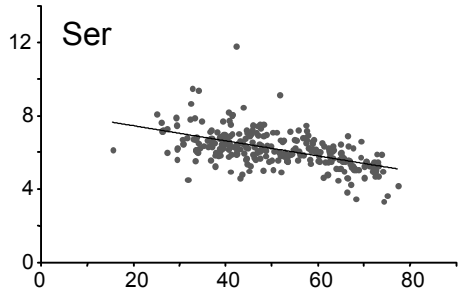


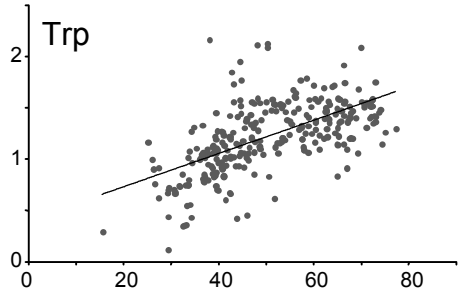
G+C global et contenu en AA

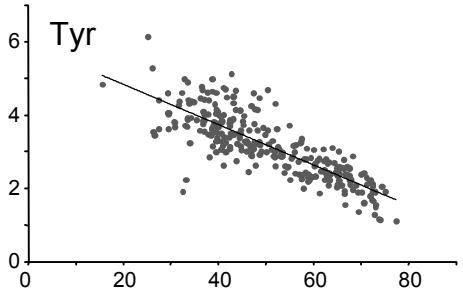


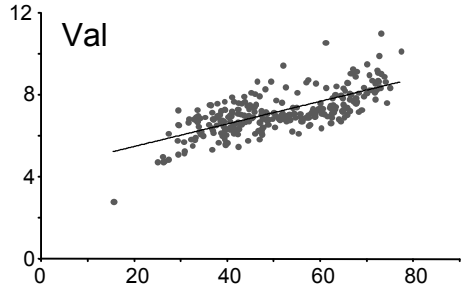












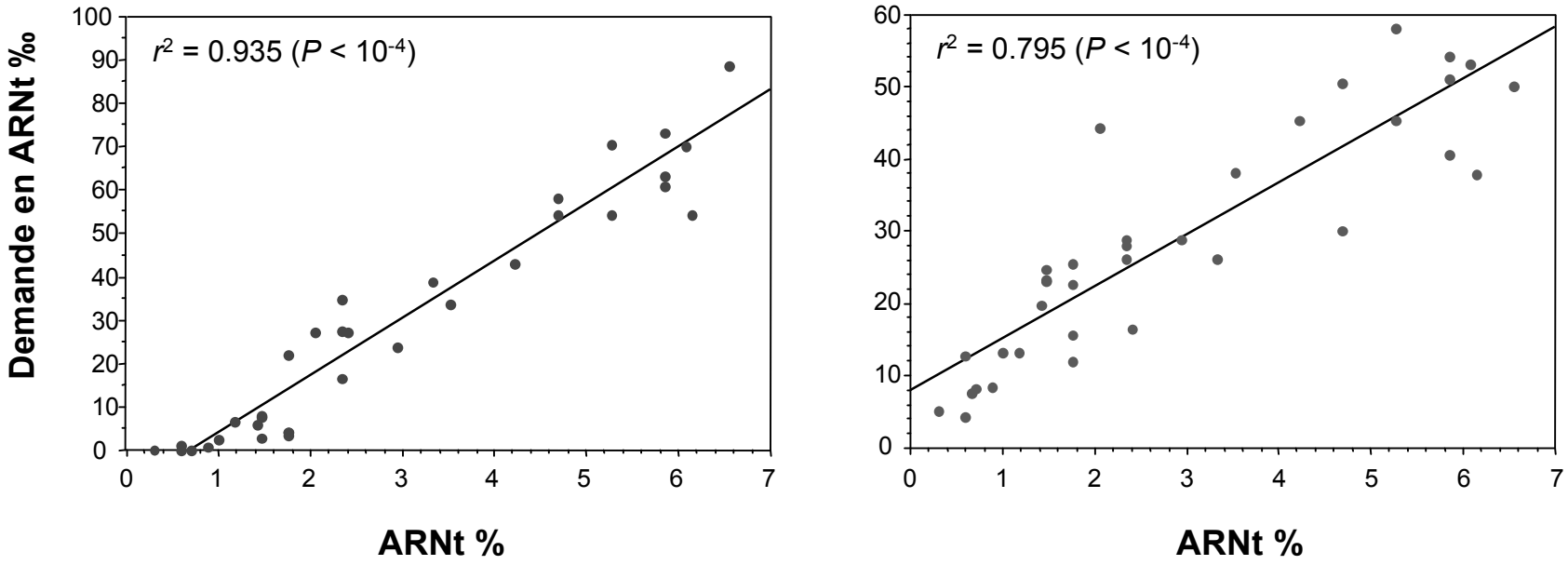
Quatre AA non corrélés au G+C% : Cys, Gln, Met, Thr

Fréquences des ARNt

AA	Codon	Anti-codon	ARNt %	AA	Codon	Anti-codon	ARNt %
Arg	CGH	ICG	5.25	Val	GUY	GAC	2.33
	CGG	idem	0.64		GUD	UAC	6.12
	AGR	idem	0.70	Lys	AAR	UUU	5.83
Leu	CUY	GAG	1.75	Asn	AAY	QUU	3.50
	CUA	idem	0.58	Gln	CAA	UUG	1.75
	CUG	CAG	5.83		CAG	CUG	2.33
	UUR	AAA	1.46	His	CAY	QUG	2.33
Ser	UCY	GGA	2.39	Glu	GAR	UUC	5.25
	UCD	UGA	1.46	Asp	GAY	QUC	4.67
	AGY	GCU	1.46	Tyr	UAY	QUA	2.92
Thr	ACY	GGU	4.67	Cys	UGY	GCA	1.17
	ACR	UGU	1.40	Phe	UUY	GAA	2.04
Pro	CCY	GGG	0.99	Ile	AUY	GAU	5.83
	CCR	UGG	3.32		AUA	NAU	0.29
Ala	GCY	GGC	4.20	Met	AUG	CAU	1.75
	GCD	UGC	6.06	Trp	UGG	CCA	1.75
Gly	GGY	GCC	6.53				
	GGA	UCC	0.88				
	GGG	CCC	0.58				

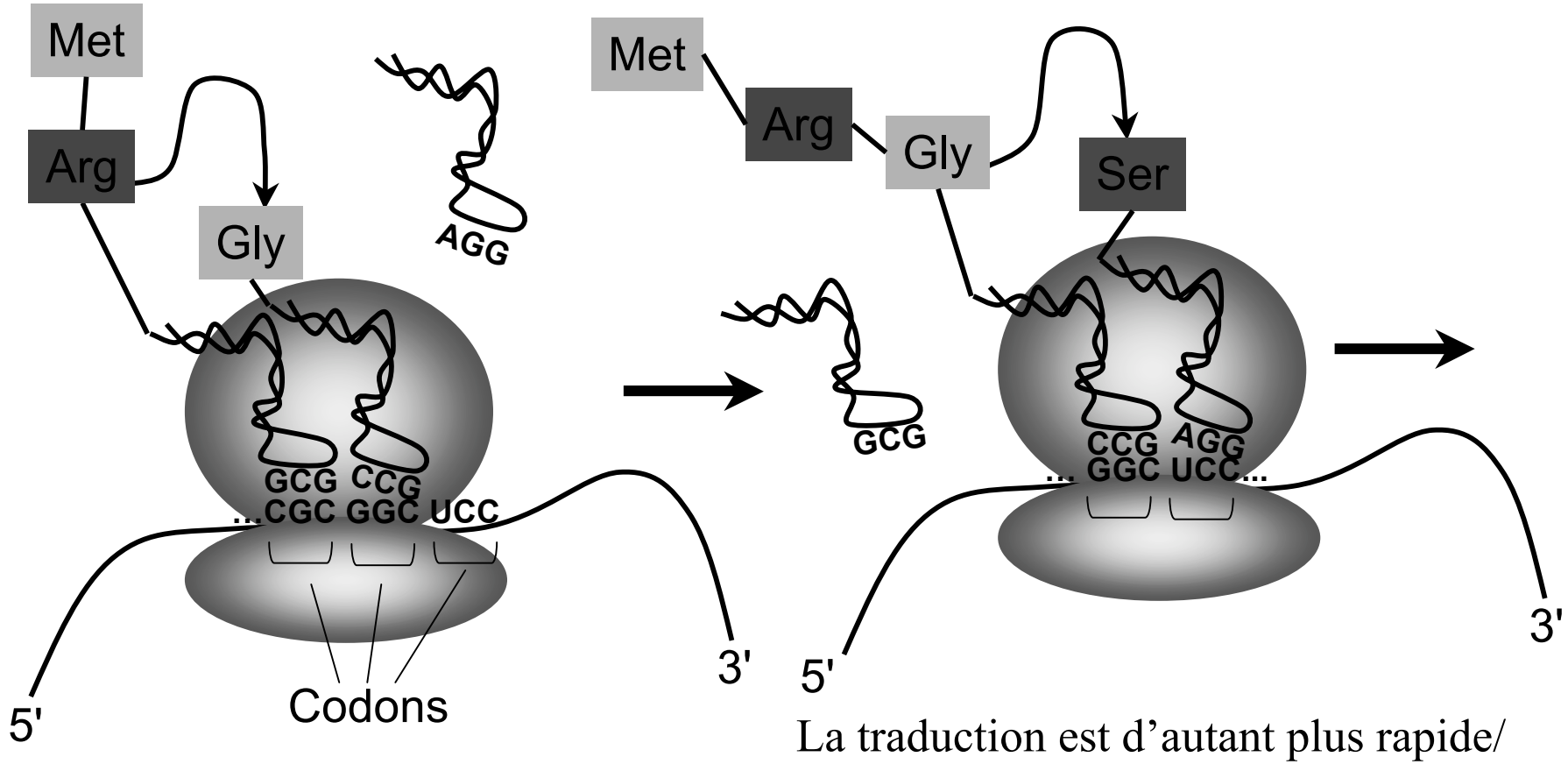
Q = Quosine, I = Inosine, R = {A, G}, Y = {C, U}, D = {A, G, U}, H = {A, C, U}

Demande en ARNt



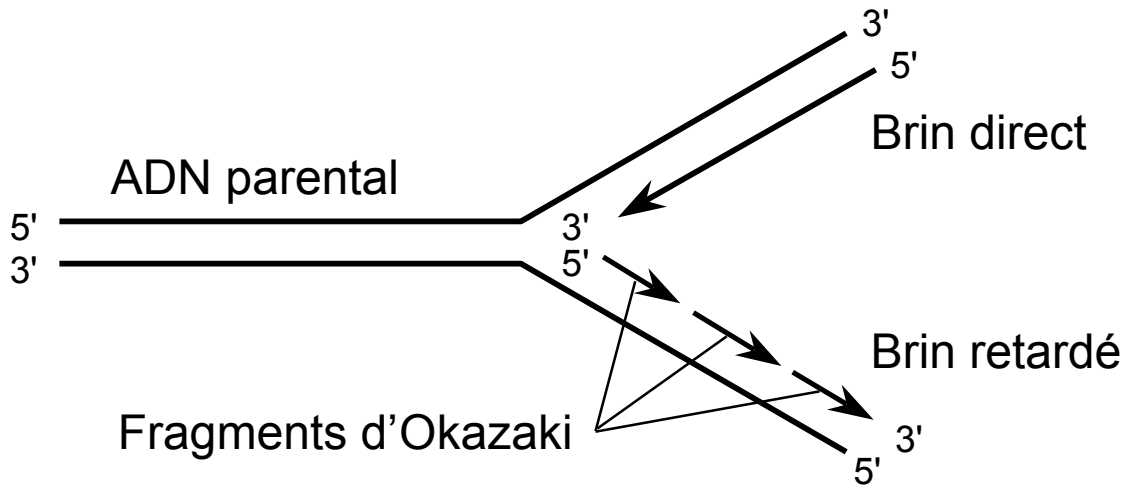
- La demande au niveau des gènes et la fréquence relative en ARNt sont corrélées positivement :
  - Cette corrélation est d'autant plus forte que les gènes sont hautement exprimés.

Sélection traductionnelle



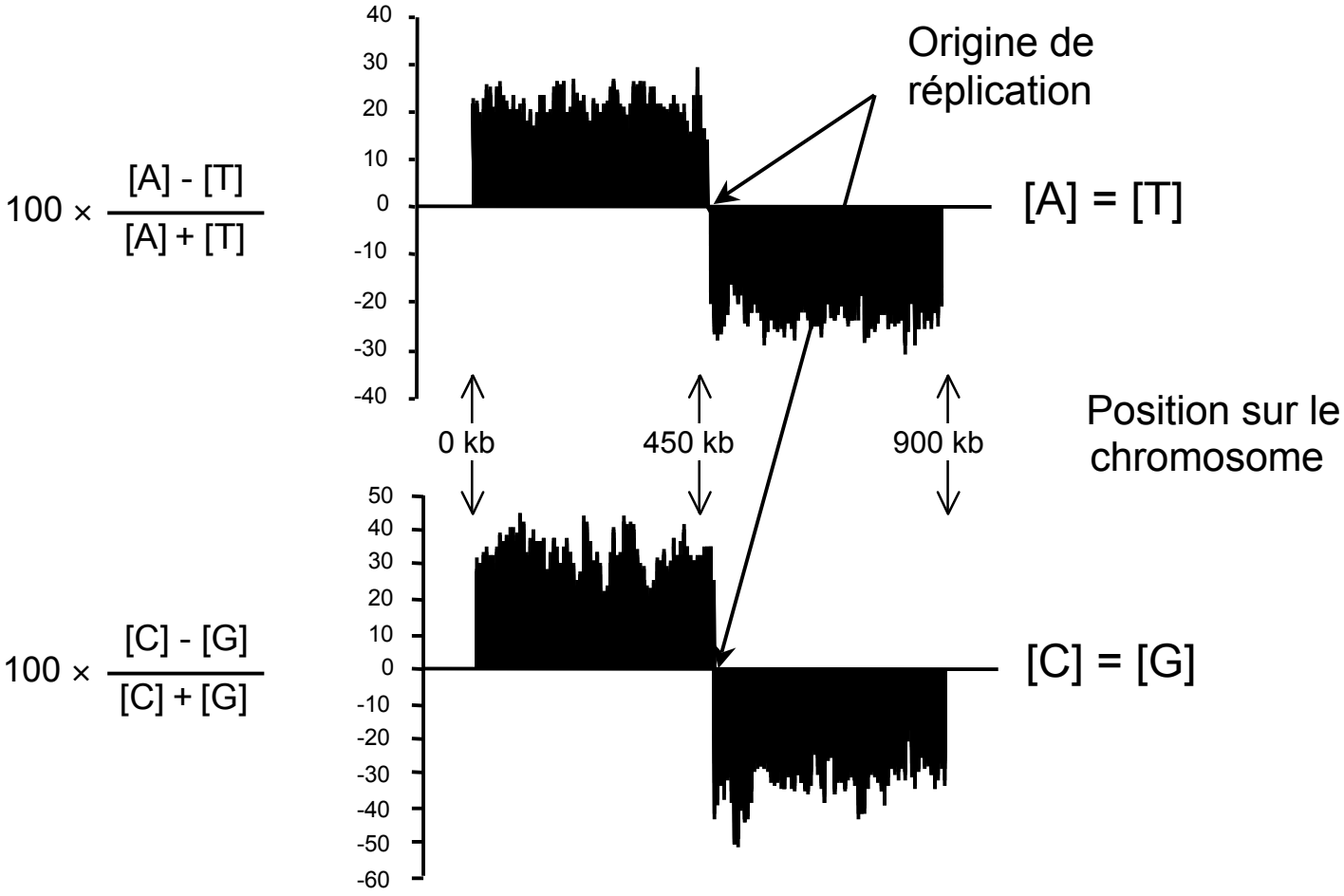
La traduction est d'autant plus rapide/ efficace que les codons utilisés correspondent à des ARNt abondants

Différence entre les deux brins



- Le brin direct et le brin retardé sont répliqués de façon différente :
  - Il existe une asymétrie de composition en base entre les deux brins chez la plupart des bactéries.
  - Cette asymétrie est très marquée chez certaines espèces.

Asymétrie chez *B. burgdorferi*



Contenu en G+C3

- Contenu en G+C3 d'un gène :
  - Pourcentage en bases G+C en position III des codons :
    - Valeur brute (G+C3).
    - Centrage par la moyenne (G+C3c).
  - La position III est le plus souvent conservatrice en cas de mutation ponctuelle :
    - Pas de changement de l'acide aminé codé dans 55 % des cas (composition uniforme).
    - La plasticité joue essentiellement sur cette position.

L'indice CAI

- Le *Codon Adaptation Index* (CAI) est calculé à partir de la formule :

$$\ln(\mathrm{CAI}) = \frac{1}{n} \sum_{i=1}^{59} n_i \ln w_i$$

où n est le nombre de codons dégénérés, n_i le nombre de codons i, et w_i le « facteur d'adaptation » du codon i.

Calcul du facteur d'adaptation

- Valeur donnée par le rapport :

$$w_i = x_i / x_{imax}$$

où x_i est la fréquence du codon i et x_{imax} est la fréquence du codon synonyme majeur pour un acide aminé donné :

  - Les valeurs de x_i et x_{imax} sont obtenues à partir d'un ensemble de gènes de référence.

Arg	AGA	5	$w_{AGA} = 5/626 = 0.0080$
	AGG	0	$w_{AGG} = 0.5/626 = 0.0008$
	CGA	2	$w_{CGA} = 2/626 = 0.0032$
	CGC	268	$w_{CGC} = 268/626 = 0.4281$
	CGG	3	$w_{CGG} = 3/626 = 0.0048$
	CGU	626	$w_{CGU} = 626/626 = 1.0000$

Choix des gènes de référence

- On utilise un ensemble de référence pour chaque espèce étudiée.
- Les gènes utilisés sont généralement hautement exprimés :
  - Fortes valeurs de CAI :
    - Gènes hautement exprimés.
  - Faibles valeurs de CAI :
    - Gènes faiblement exprimés.
    - Gènes transférés horizontalement.

χ^2 d'usage du code

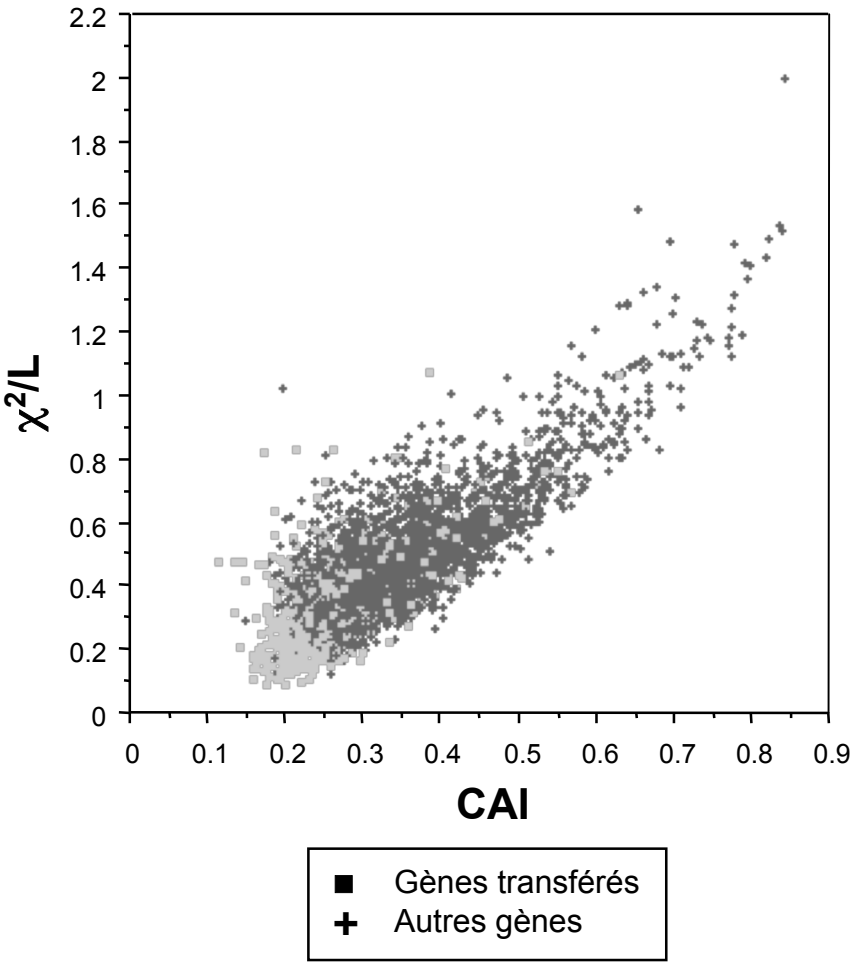
- Mesure de l'utilisation des codons synonymes par rapport à un usage aléatoire.

$$\chi^2 = \sum_{i=1}^{20} \sum_{j=1}^{s_j} \frac{(m_{ij} - m_{i\bullet}/s_j)^2}{m_{i\bullet}/s_j}$$

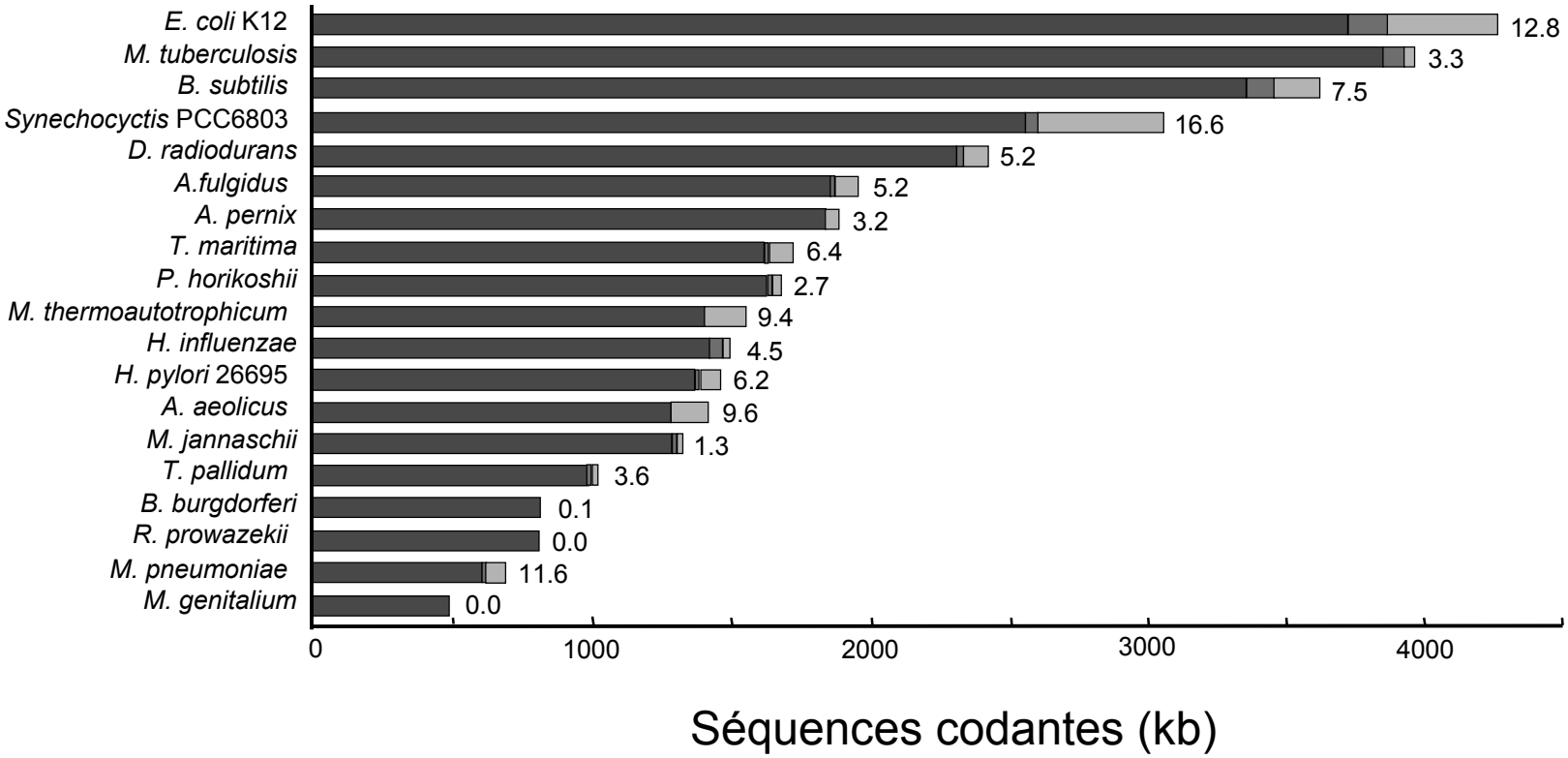
où m_{ij} est le nombre de codons j codants pour l'acide aminé i, $m_{i\bullet}$ le nombre total de codons codants pour i, et s_j le nombre de codons synonymes pour i.

Lawrence et Ochman (1998)

- Utilisation du CAI et d'une mesure du χ^2 pondérée par la longueur des gènes.
- Proximité de structures impliqués dans les transferts :
  - Élements transposables (séquences IS).
  - Régions recombinantes (sites *chi*).



Autres espèces

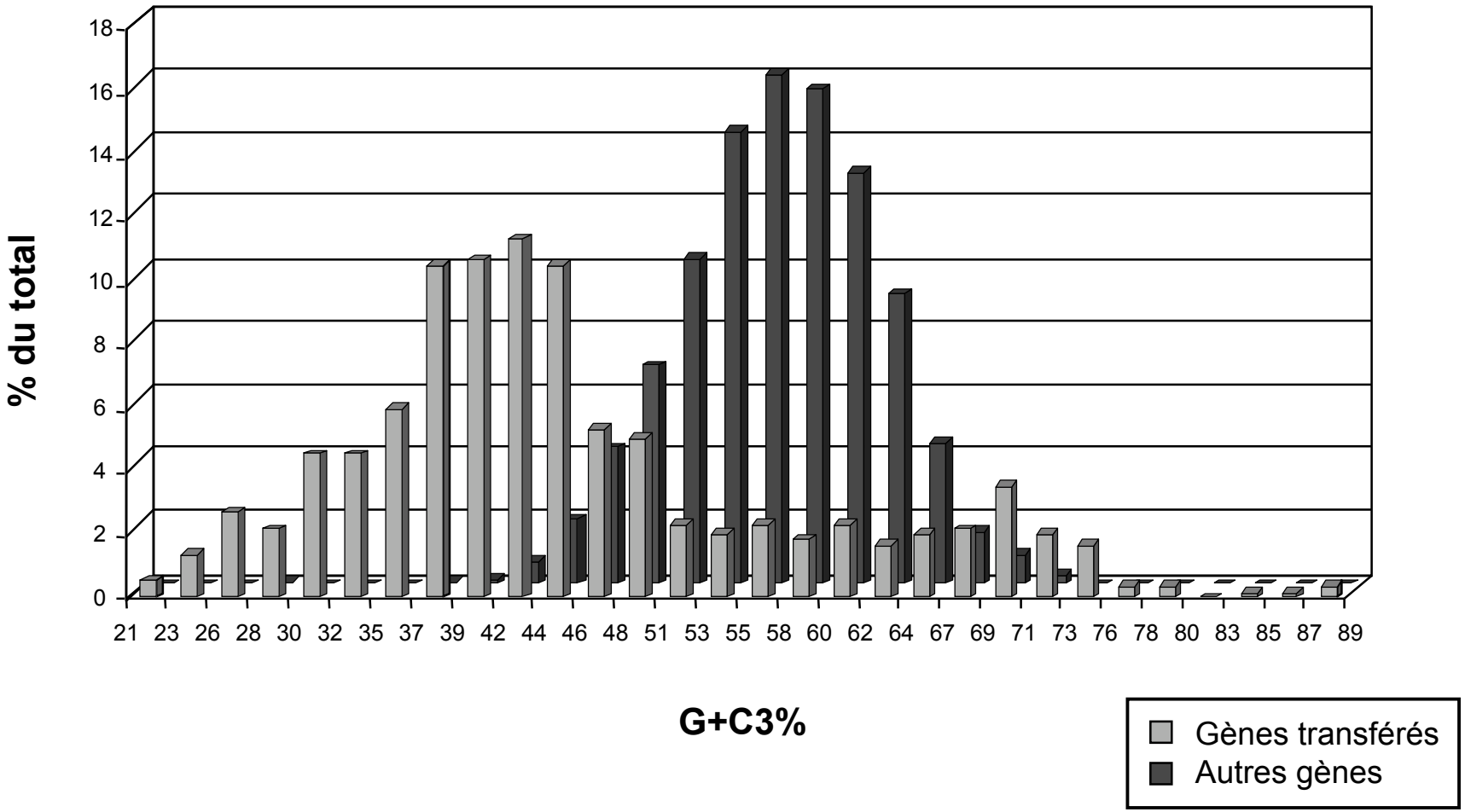


- Gènes « natifs » présents dans le génome
- Gènes transférés associés à des éléments mobiles
- Gènes transférés non associés à des éléments mobiles

Résultats chez *E. coli*

- 755 gènes sur 4286 (17,6 %) sont prédits comme transférés chez *E. coli*.
- 67 % des séquences IS du génome sont associées à ces gènes :
  - Ces séquences IS constituent fréquemment des frontières entre zones endogènes et exogènes.
- 36 % des gènes détectés se situent au voisinage du terminus de réplication (24 % du chromomosome).

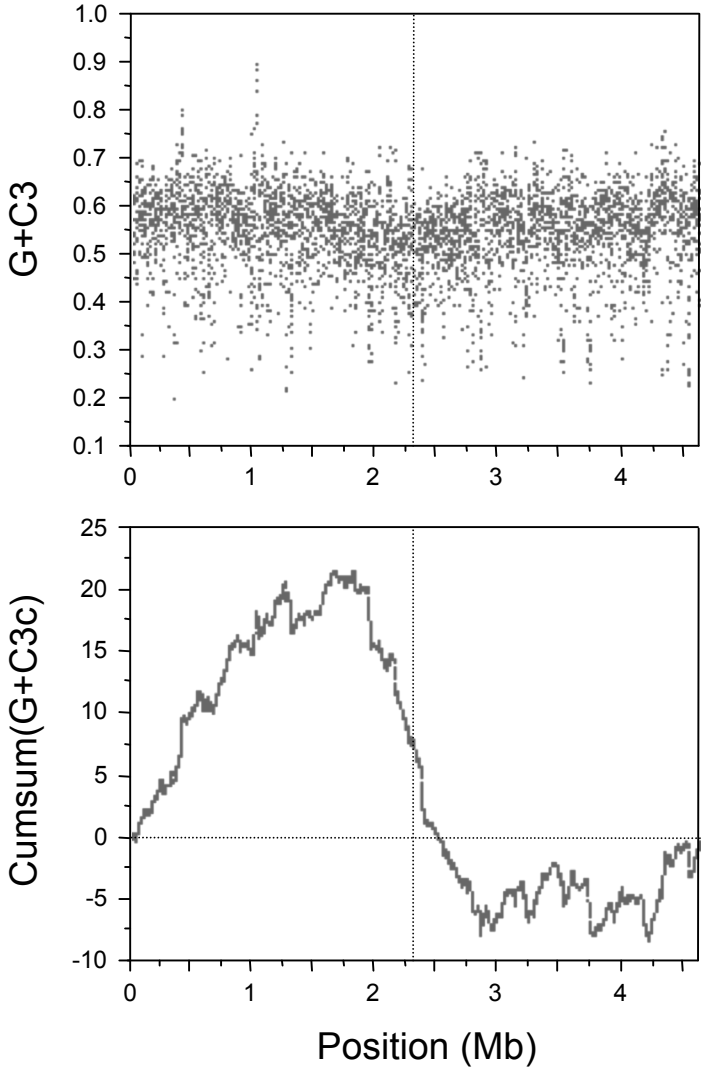
G+C3 des gènes



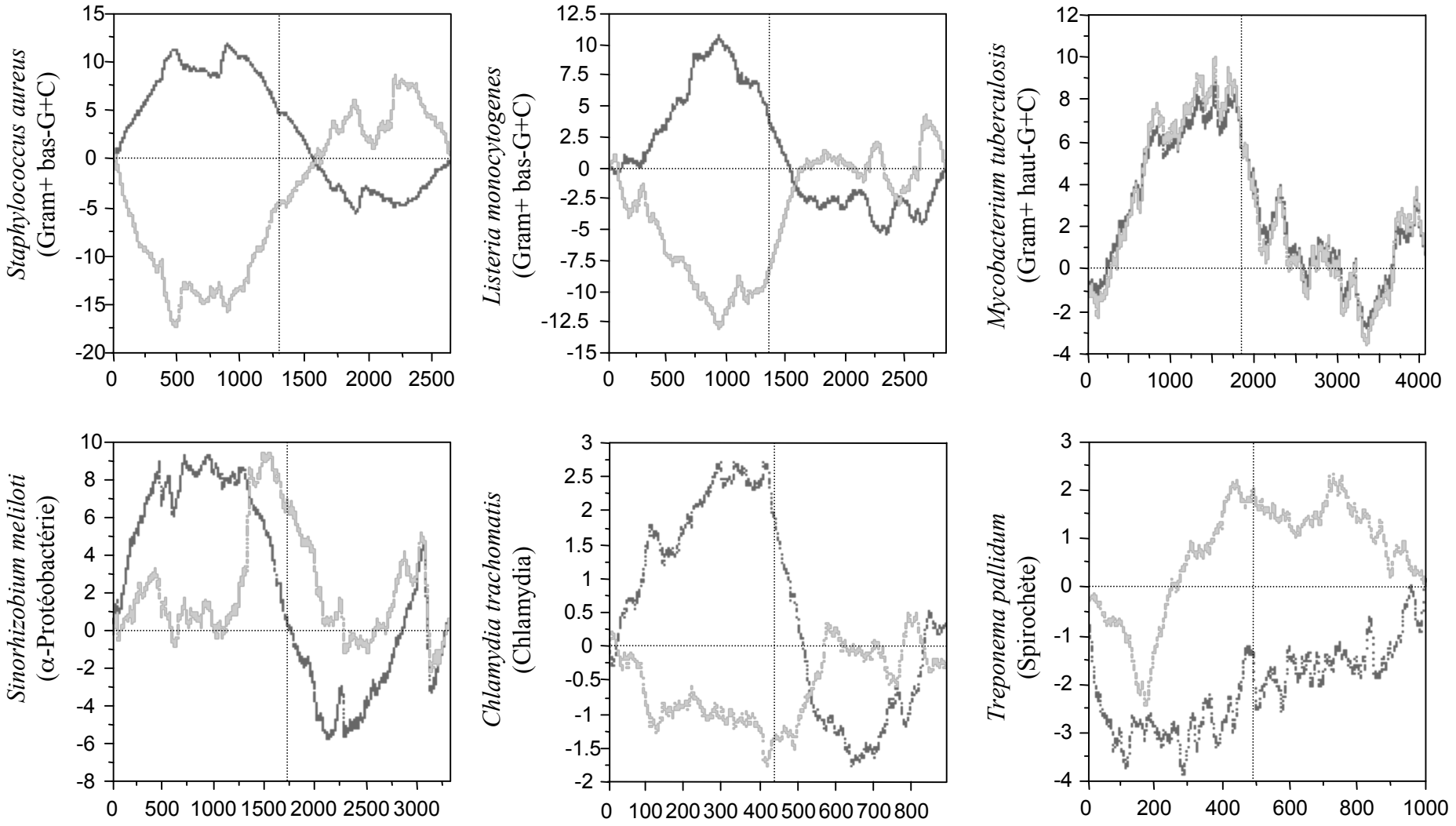
Effet de la localisation

La variation des valeurs « brutes » du G+C3 le long du chromosome d'*E. coli* n'est pas significative

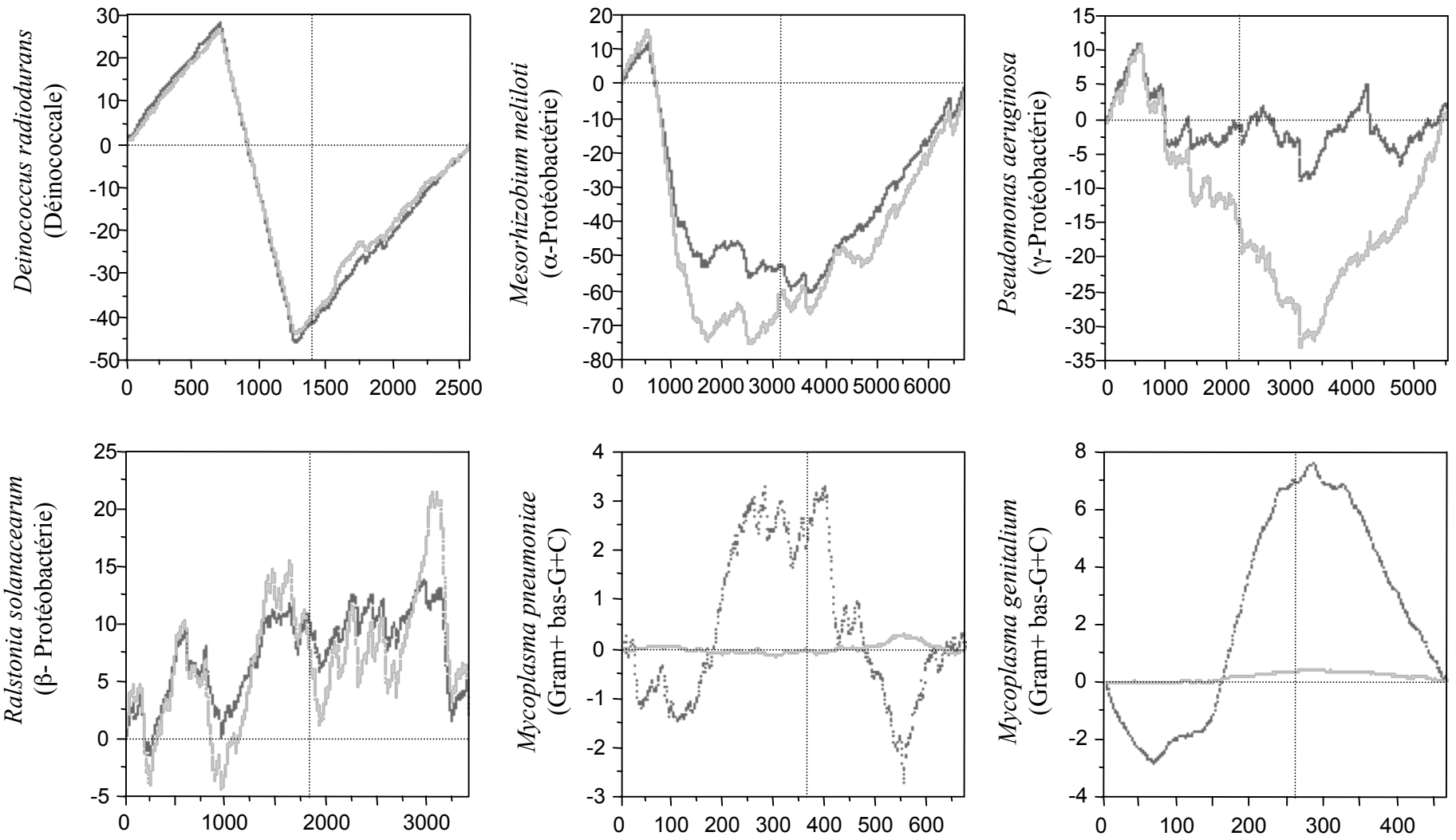
L'utilisation des valeurs centrées / cumulées (G+C3c) révèle une structuration du génome



Espèces similaires à *E. coli*

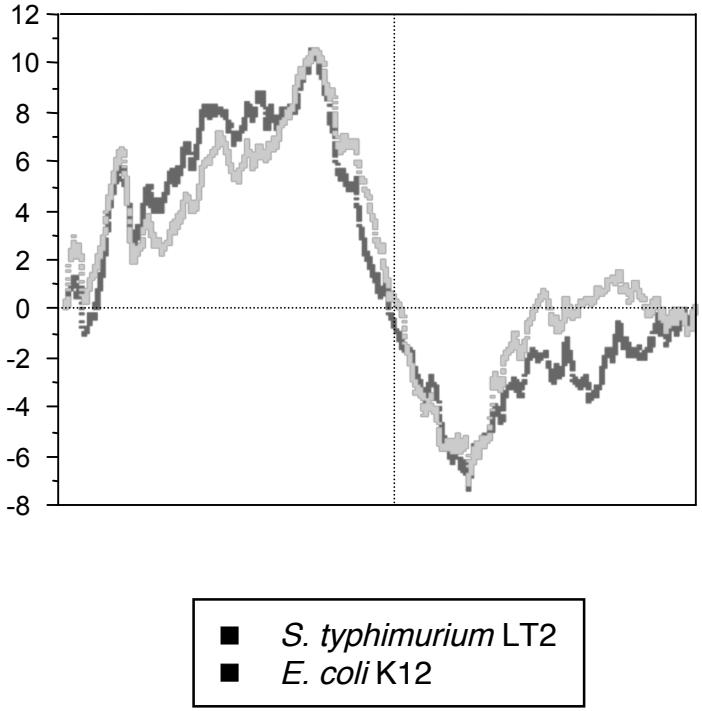


Cas particuliers

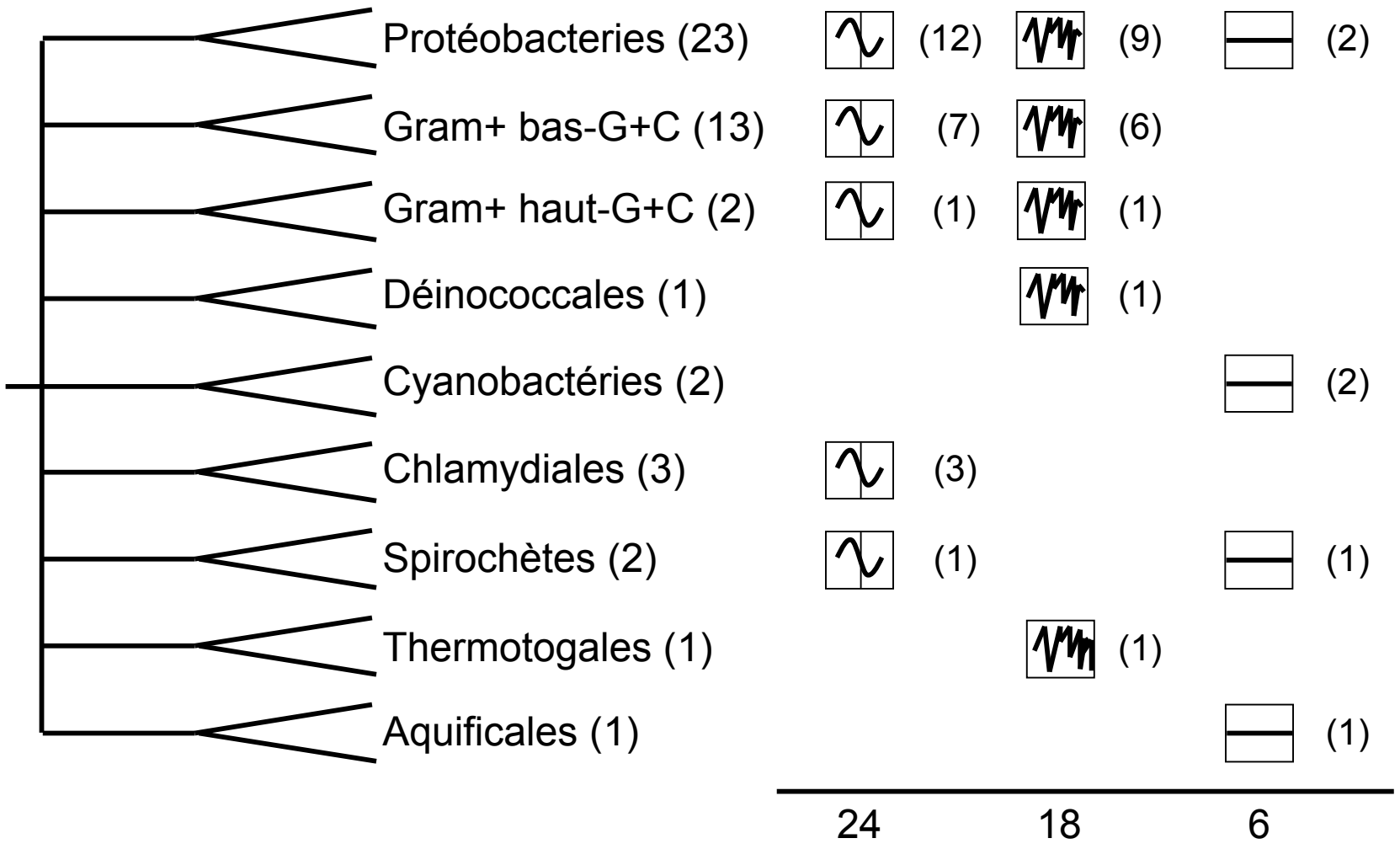


Escherichia / Salmonella

- Utilisation des gènes conservés entre *E. coli* et *S. typhimurium*:
  - Le motif est retrouvé dans les deux espèces :
    - Pourquoi ce bais aurait-il été conservé dans le cas d'un transfert massif à l'ancêtre commun ?
    - Enrichissement intrinsèque au lieu de transferts.

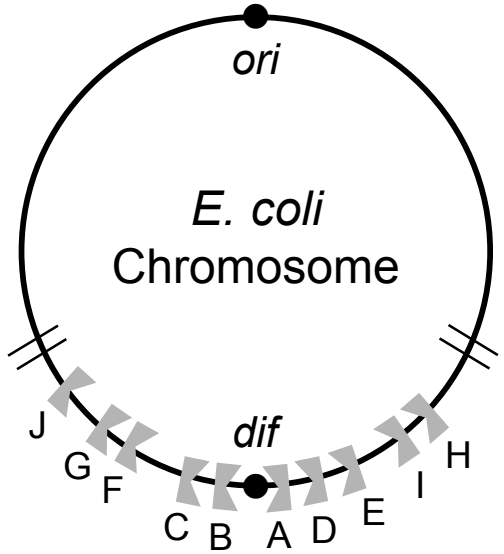


Répartition phylogénétique



Le terminus d'*E. coli*

- La région du terminus contient dix sites *ter* (A-J) :
  - Combinaison à la protéine Tus :
    - Inhibition de l'action des hélicases de manière polaire.
    - Les deux fourches de réplication se rejoignent au niveau du site *dif*.
    - Résolution des dimères.



Localisation et polarité des dix sites *ter* d'*E. coli*

Pourquoi un tel motif ?

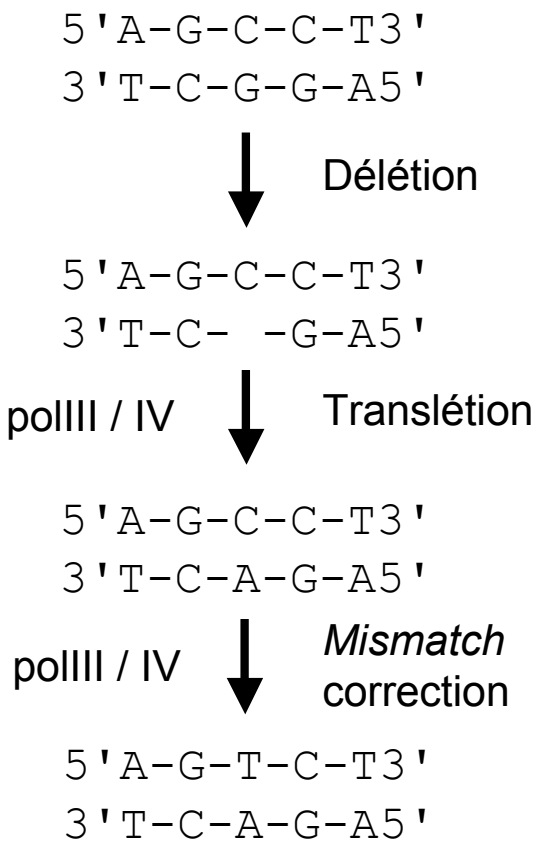
- Les gènes transférés s'insèrent plus fréquemment au niveau du terminus :
  - Ces gènes seraient systématiquement plus riches en bases A+T que l'hôte.
- Existence d'un phénomène intrinséque, indépendant des transferts :
  - Biais mutationnel déplacé vers A+T dans cette région.

Réparation des lésions

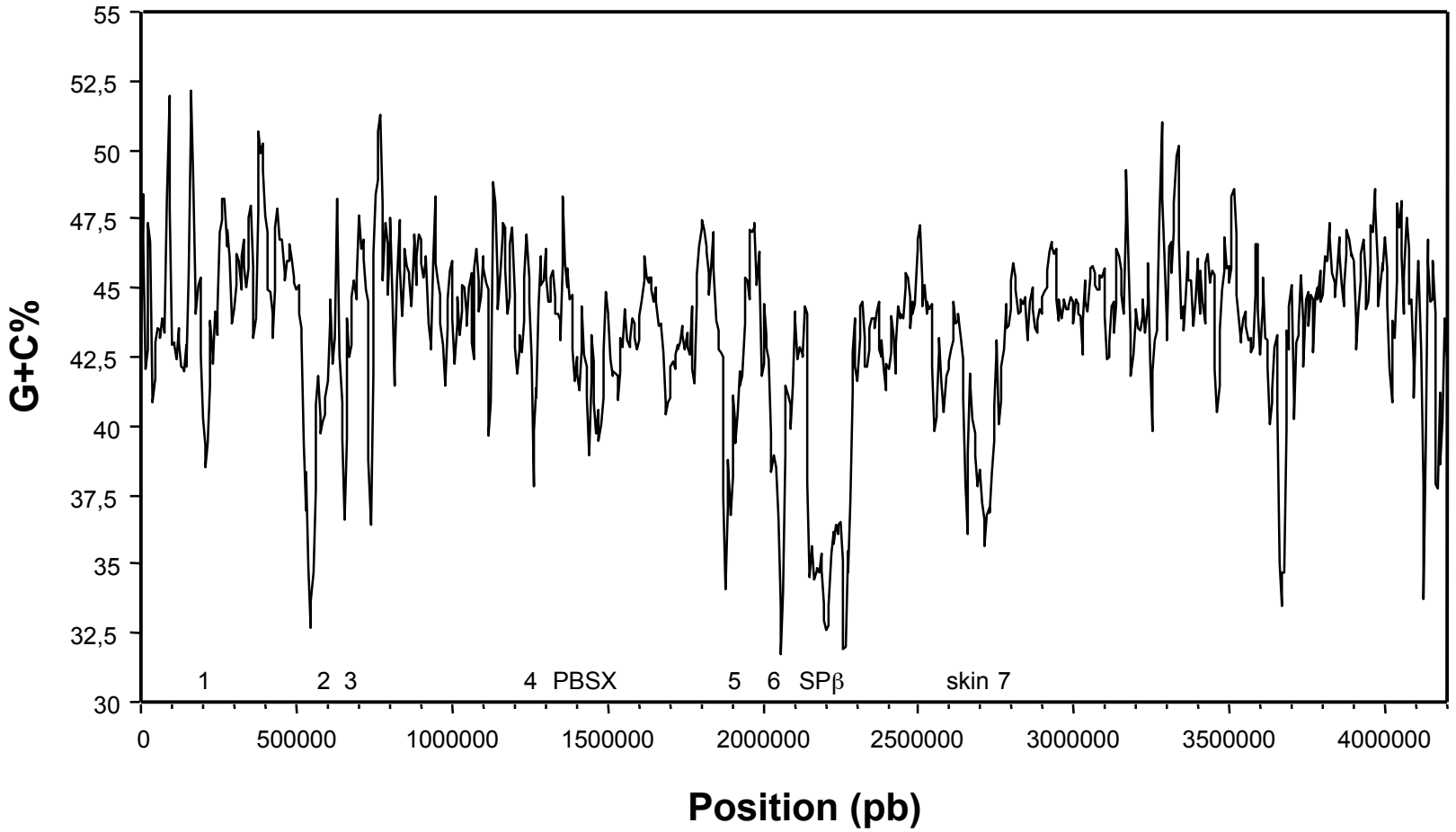
- Quand le complexe atteint une lésion, la réplication s'arrête :
  - Recul de la fourche.
  - Formation d'une jonction de Holliday sous l'action des hélicases RecG et PriA :
    - Réparation par recombinaison homologue.
- Dans la région du terminus, les complexes *ter*/Tus inhibent l'action des hélicases !
  - Réparation par le mécanisme de translésion.

La translétion

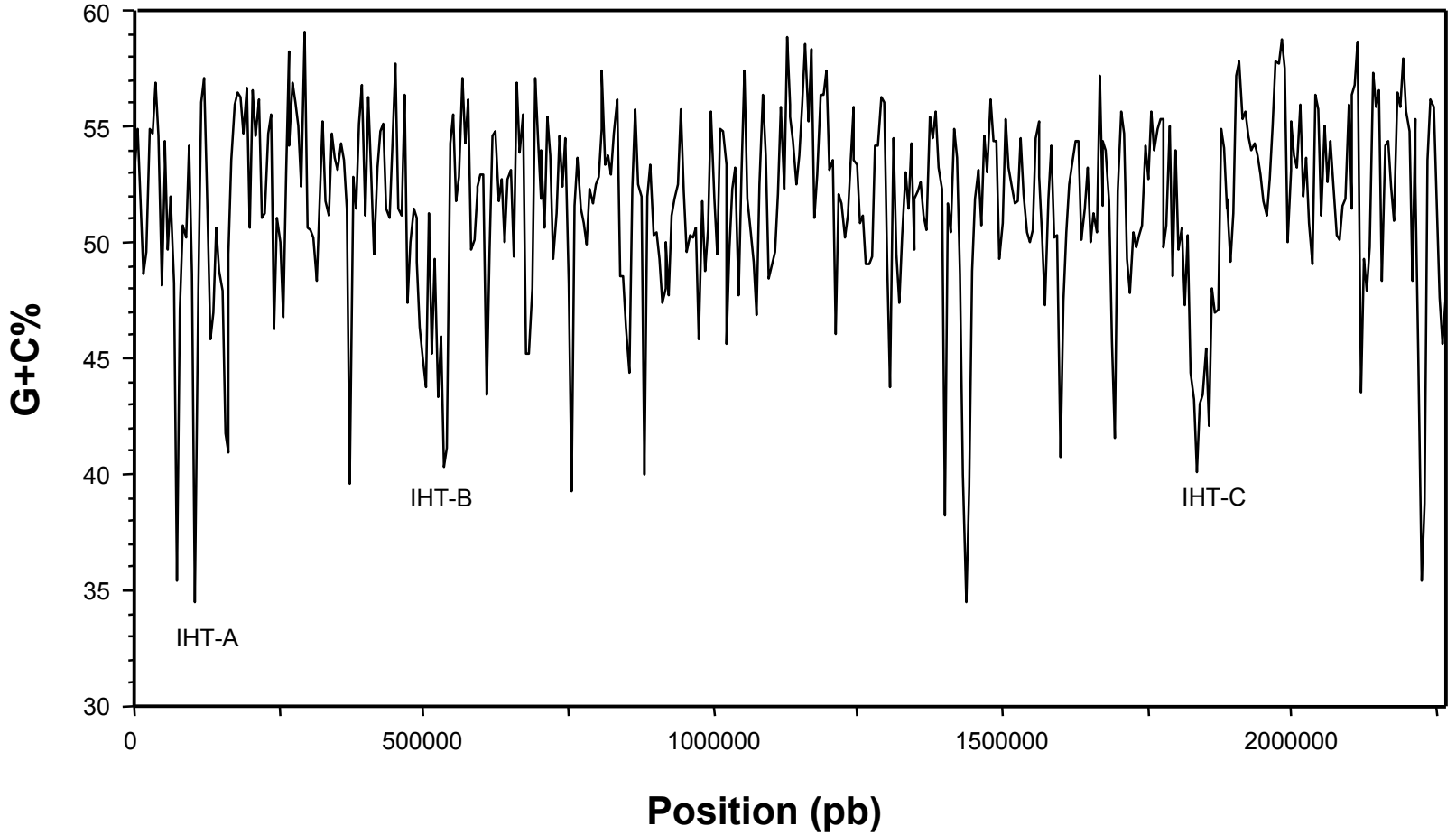
- Réalisée par l'intermédiaire des polymérases SOS (polIII et polIV) :
  - Introduction systématique de dAMP (*A-rule*) aux sites abasiques :
    - Enrichissement en A+T de régions n'utilisant pas la réparation par recombinaison.



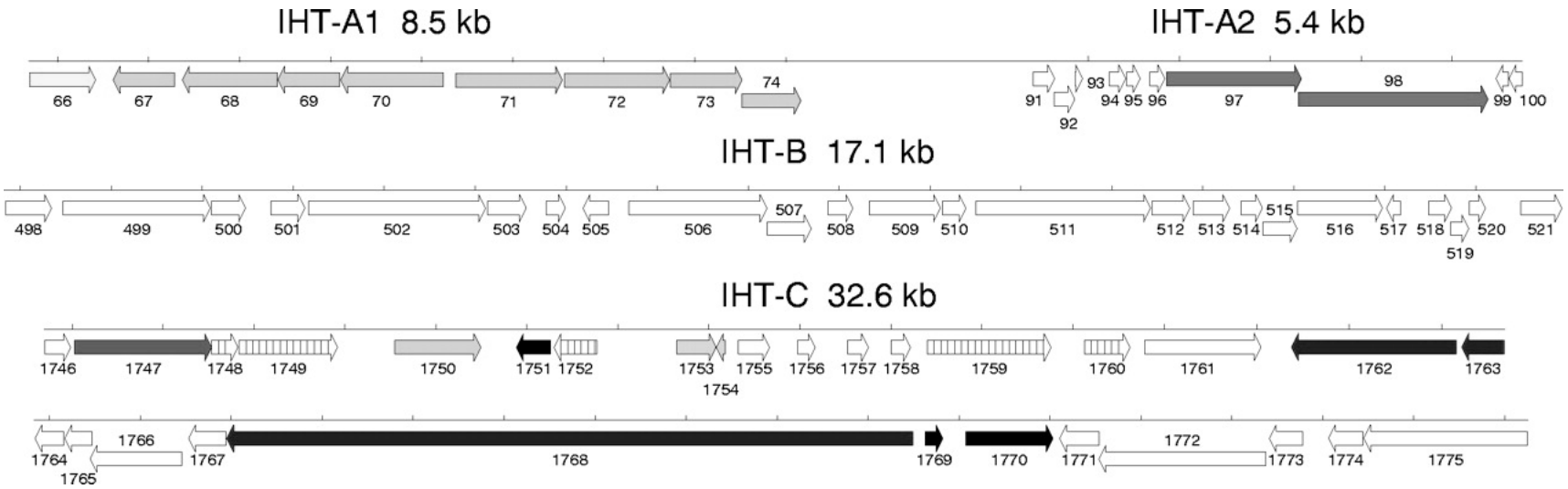
Profil en G+C chez *B. subtilis*



Profil chez *N. meningitidis*



Les ilôts de *N. meningitidis*



IHT-A1

66 : adénine rRNA méthylase, 67-70 : protéines de biosynthèse de la capsule, 71-74 : protéines d'exportation de la capsule

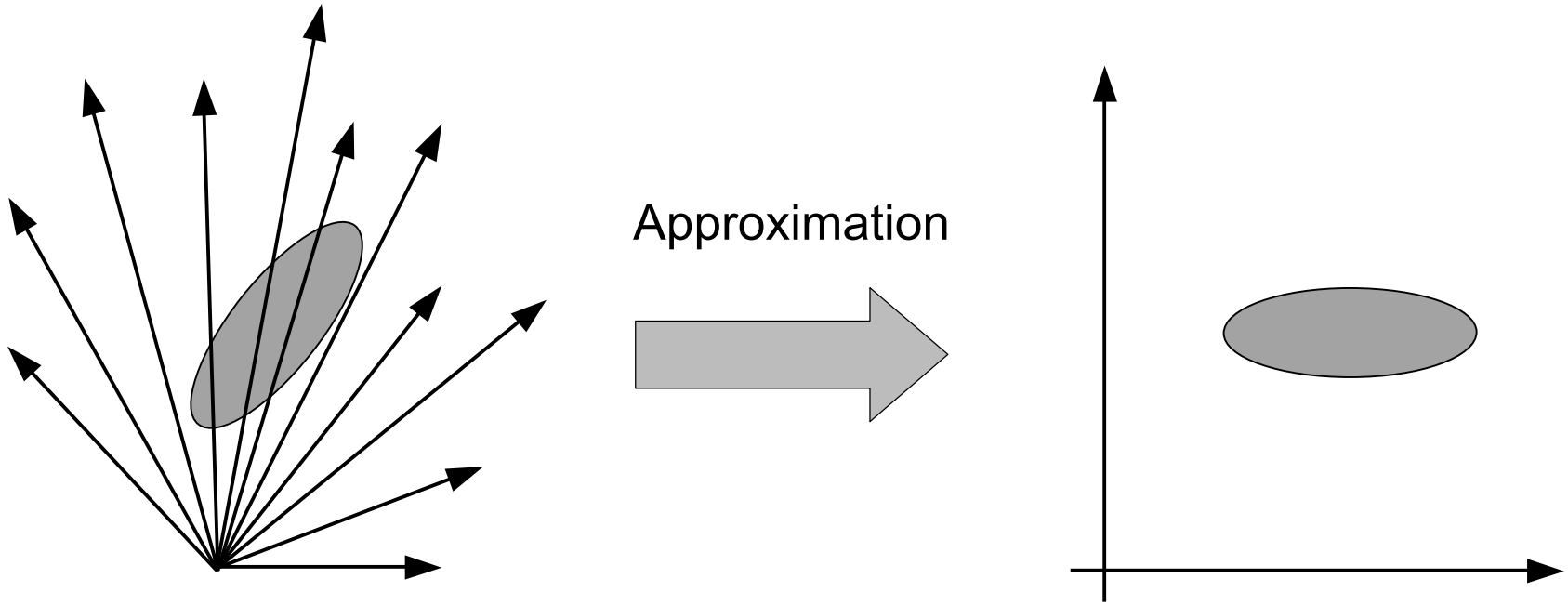
IHT-A2

97 : protéine de sécrétion, 98 : transporteur ABC

IHT-C

1747 : protéine TspB, 1750 : PivNM-2, 1751, 1769-70 : transposases, 1753-54 : protéines de phages, 1762-63, 1768 : toxines ou assimilées

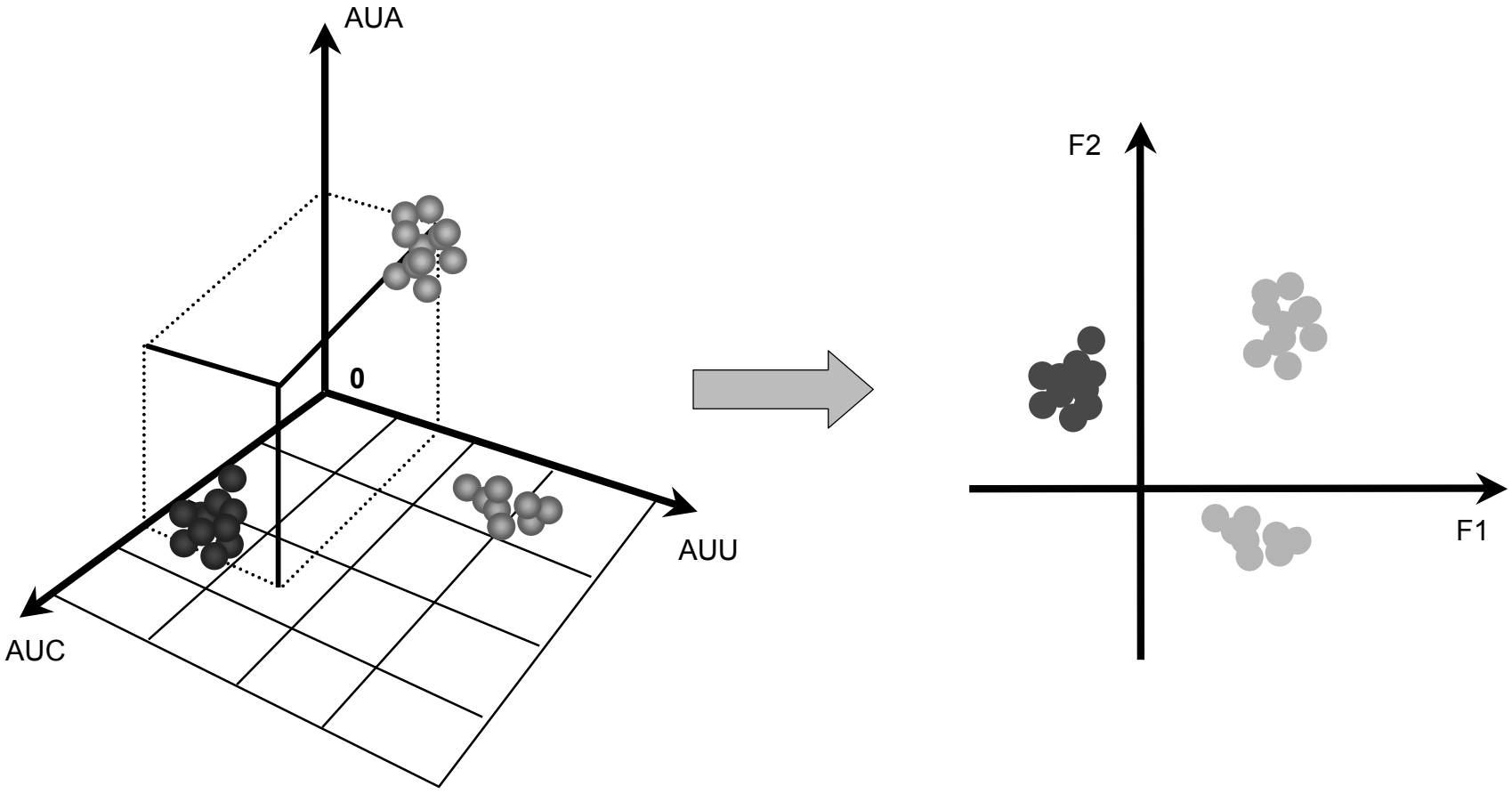
Analyses multivariées



Espace de grande dimension

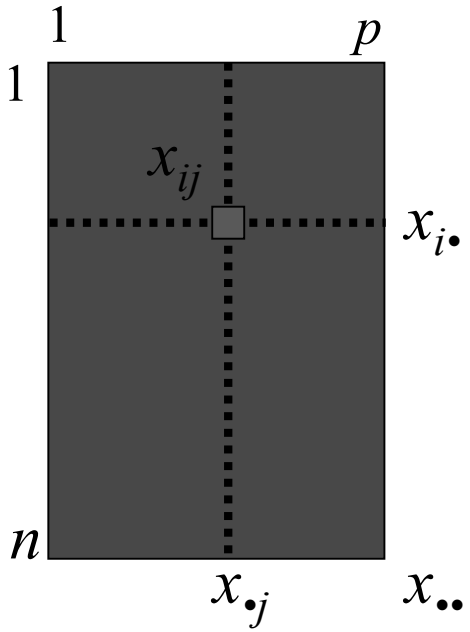
Espace à deux dimensions

Exemple de l'isoleucine



Tableaux utilisables par l'AFC

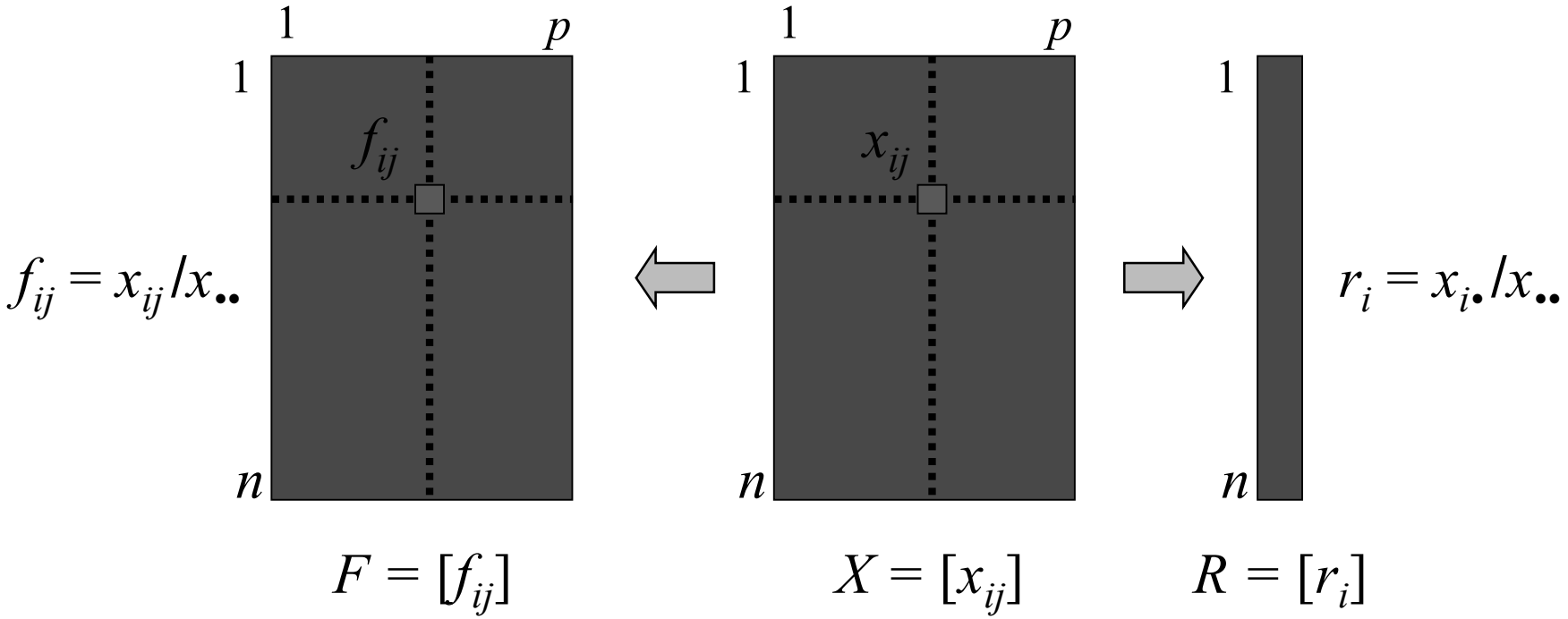
- Tableaux de contingence :
  - Tableaux ventilant une population selon des caractères croisés.
- Par extension :
  - Tous tableaux contenant des fréquences absolues (effectifs).
    - Les marges (sommes lignes et colonnes) ont une signification.



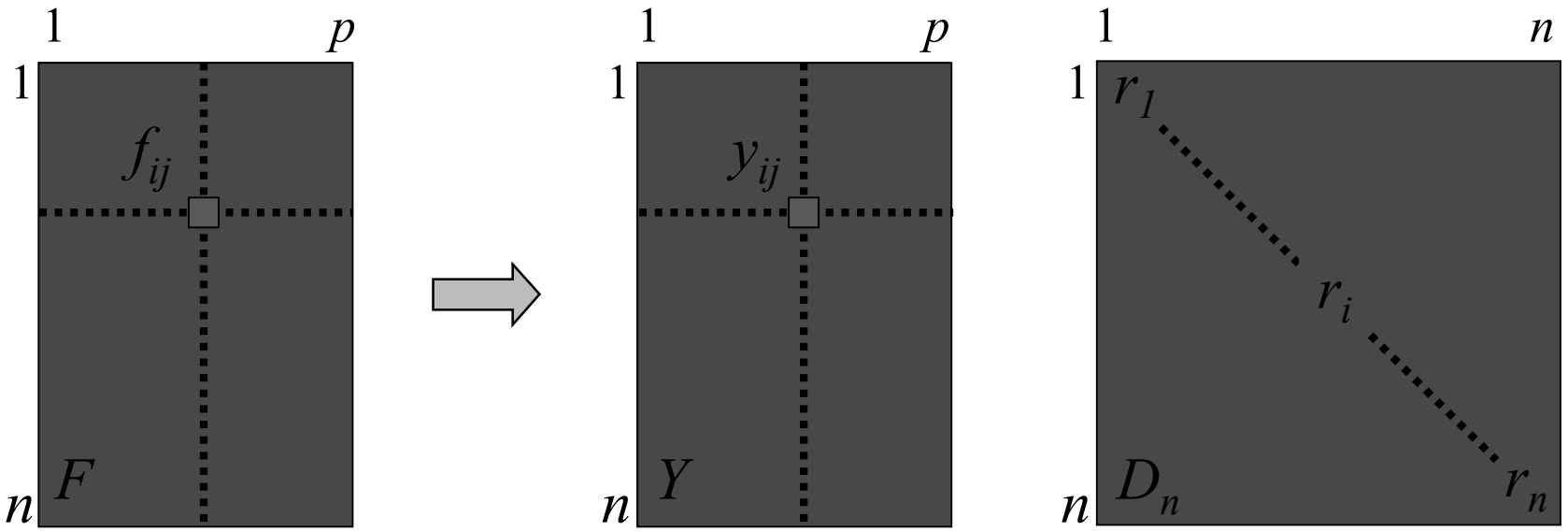
$$X = [x_{ij}]$$

$p = 61$ (codons Ter exclus)

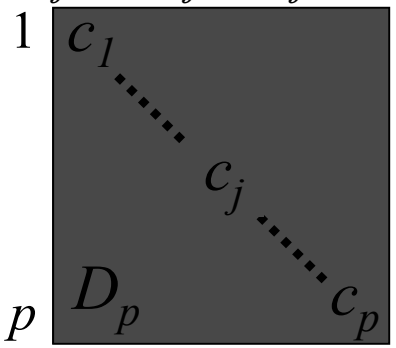
Principe de l'AFC (1)



Principe de l'AFC (2)



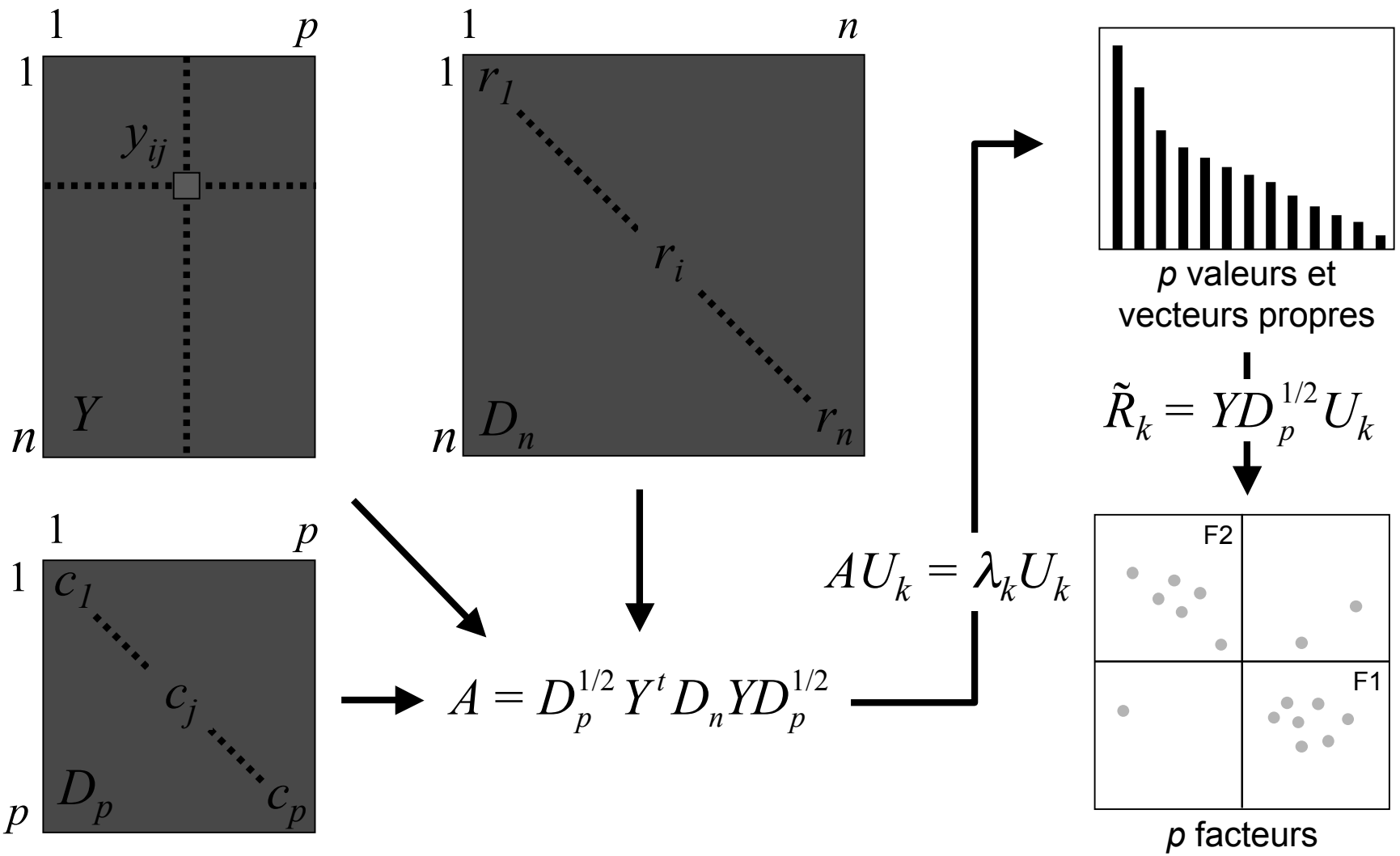
$$y_{ij} = (f_{ij} / r_i c_j) - 1$$



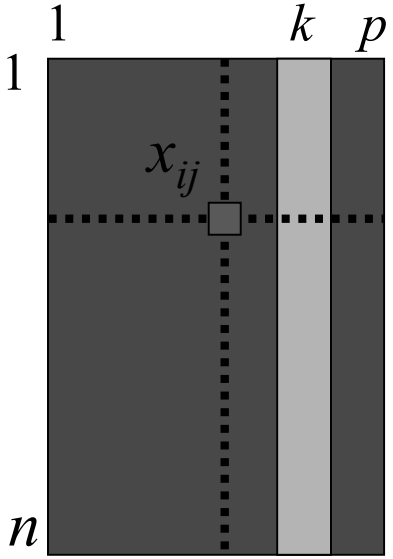
D_p : matrice diagonale construite avec les éléments de C

D_n : matrice diagonale construite avec les éléments de R

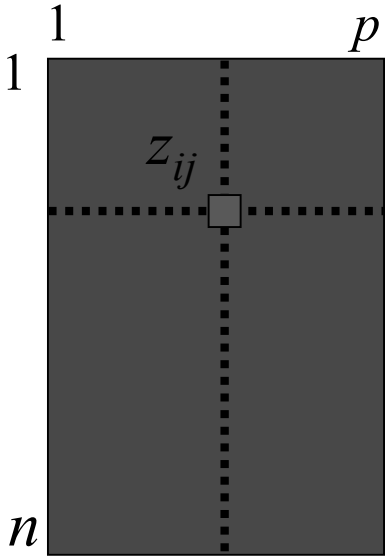
Principe de l'AFC (3)

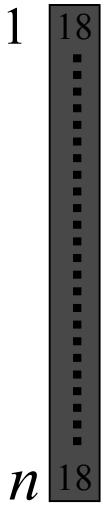


Fréquences relatives



$$z_{ij} = \frac{x_{ij}}{\sum_{j' / Cl(j') = k} x_{ij'}}$$





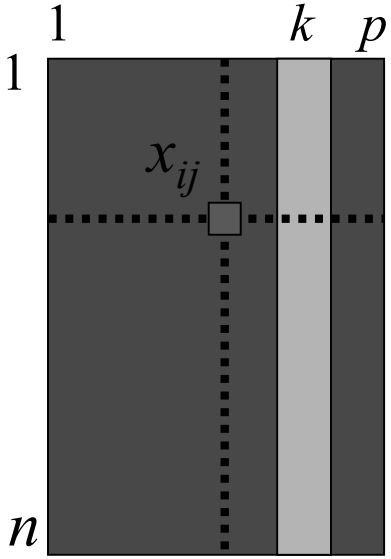
$X = [x_{ij}]$ $Z = [z_{ij}]$ $R = [r_i]$

p = 59 (codons Met et Trp exclus)

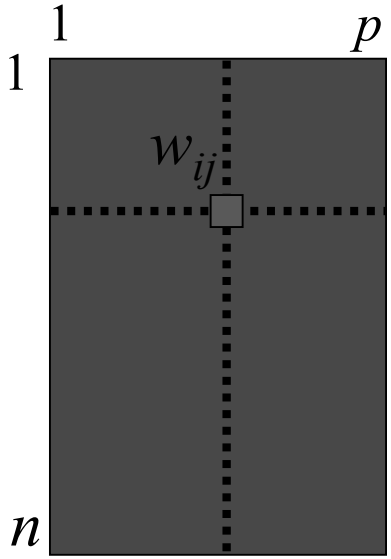
s_k = nombre de codons synonymes pour k

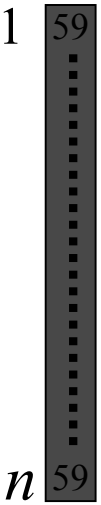
$z_{ij} = 1/s_k \ \forall j'$ si pas d'acide aminé k

Relative Synonymous Codon Usage



$$w_{ij} = \frac{x_{ij}}{\frac{1}{s_k} \sum_{j' / Cl(j') = k} x_{ij'}}$$





$X = [x_{ij}]$ $W = [w_{ij}]$ $R = [r_i]$

p = 59 (codons Met et Trp exclus)

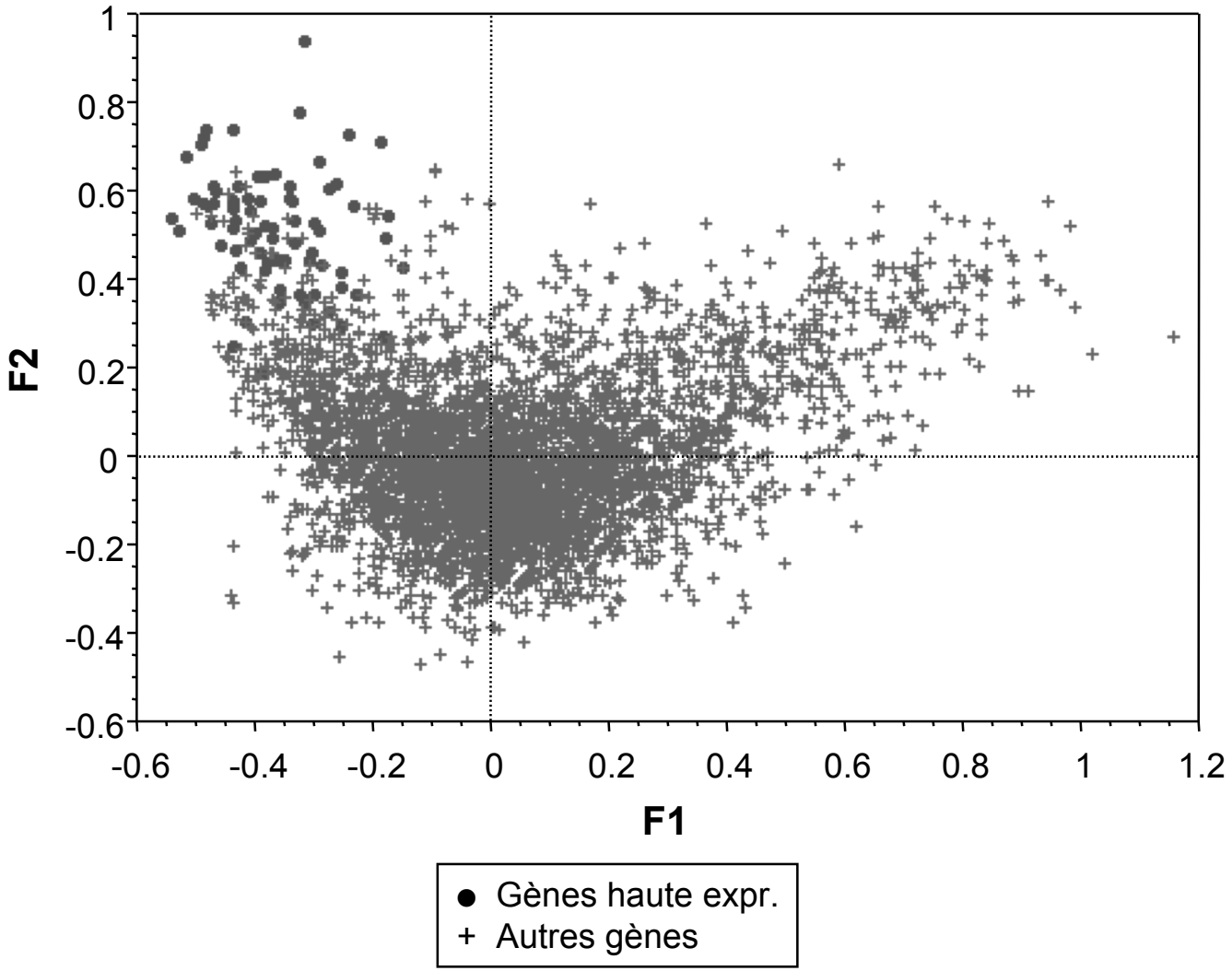
s_k = nombre de codons synonymes pour k

w_{ij} = 1 $\forall j'$ si pas d'acide aminé k

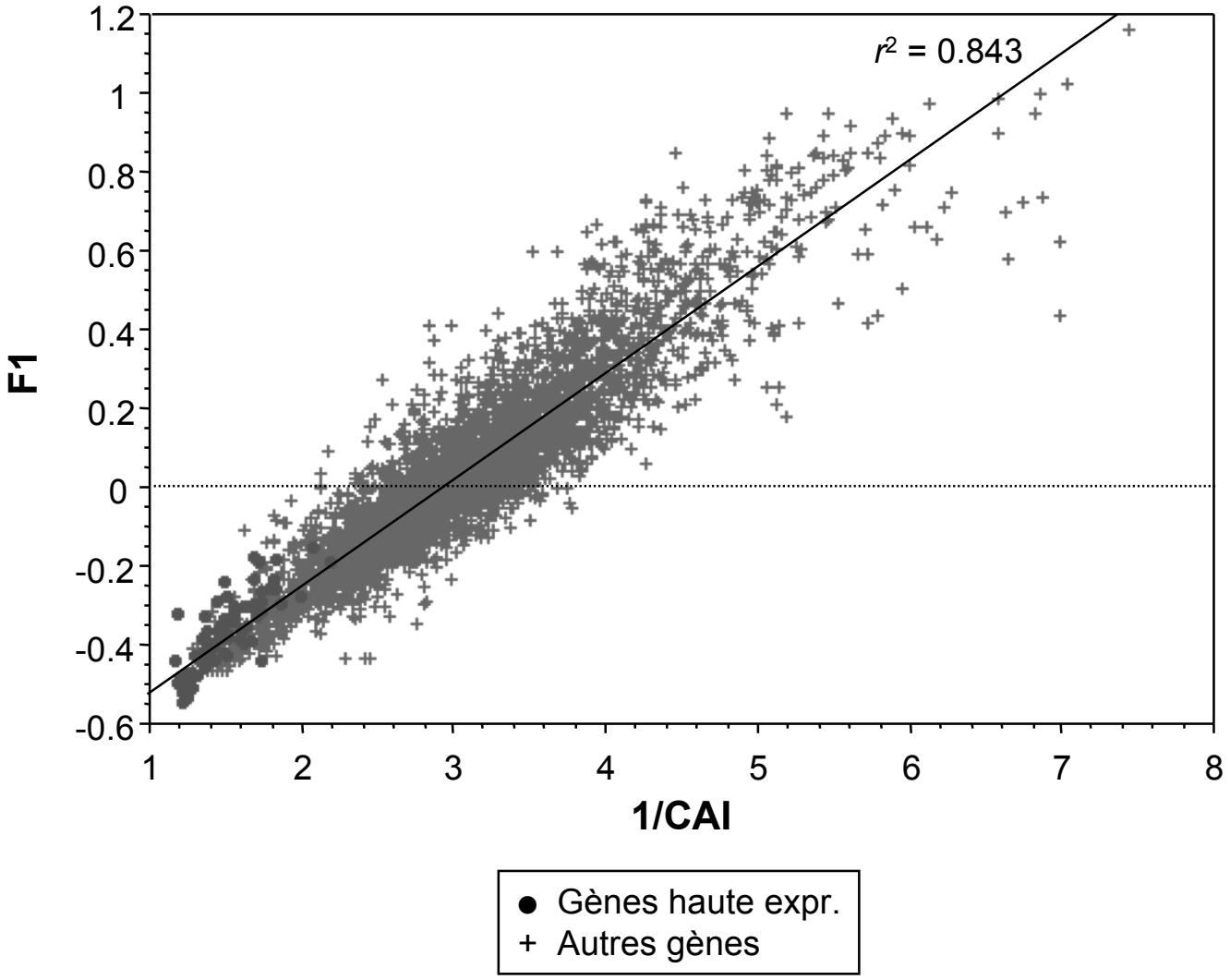
Études réalisées avec l'AFC

Espèce	Article	Tableau
Bacillus subtilis	Shields et Sharp (1987)	FA
	Sharp *et al.* (1990)	RSCU
	Perrière *et al.* (1994)	FA
	Moszer *et al.* (1995)	FR
	Moszer *et al.* (1999)	FR
Borrelia burgdorferi	McInerney (1998)	RSCU
	Lafay *et al.* (1999)	RSCU
Chlamydia trachomatis	Romero *et al.* (2000)	RSCU
Caenorhabditis elegans	Stenico *et al.* (1994)	RSCU
Drosophila melanogaster	Shields *et al.* (1988)	RSCU
Escherichia coli	Holm (1986)	FA
	Médigue *et al.* (1991)	FR
Helicobacter pylori	Lafay *et al.* (2000)	FA+RSCU
Mycoplasma genitalium	McInerney (1997)	RSCU
Pseudomonas aeruginosa	Gupta et Ghosh (2001)	RSCU
	Grocock et Sharp (2002)	RSCU
Rickettsia prowazekii	Andersson et Sharp (1996)	FA+RSCU
Thermotoga maritima	Zavala *et al.* (2002)	RSCU
Treponema pallidum	Lafay *et al.* (1999)	RSCU
Divers	Grantham *et al.* (1980ab)	FA

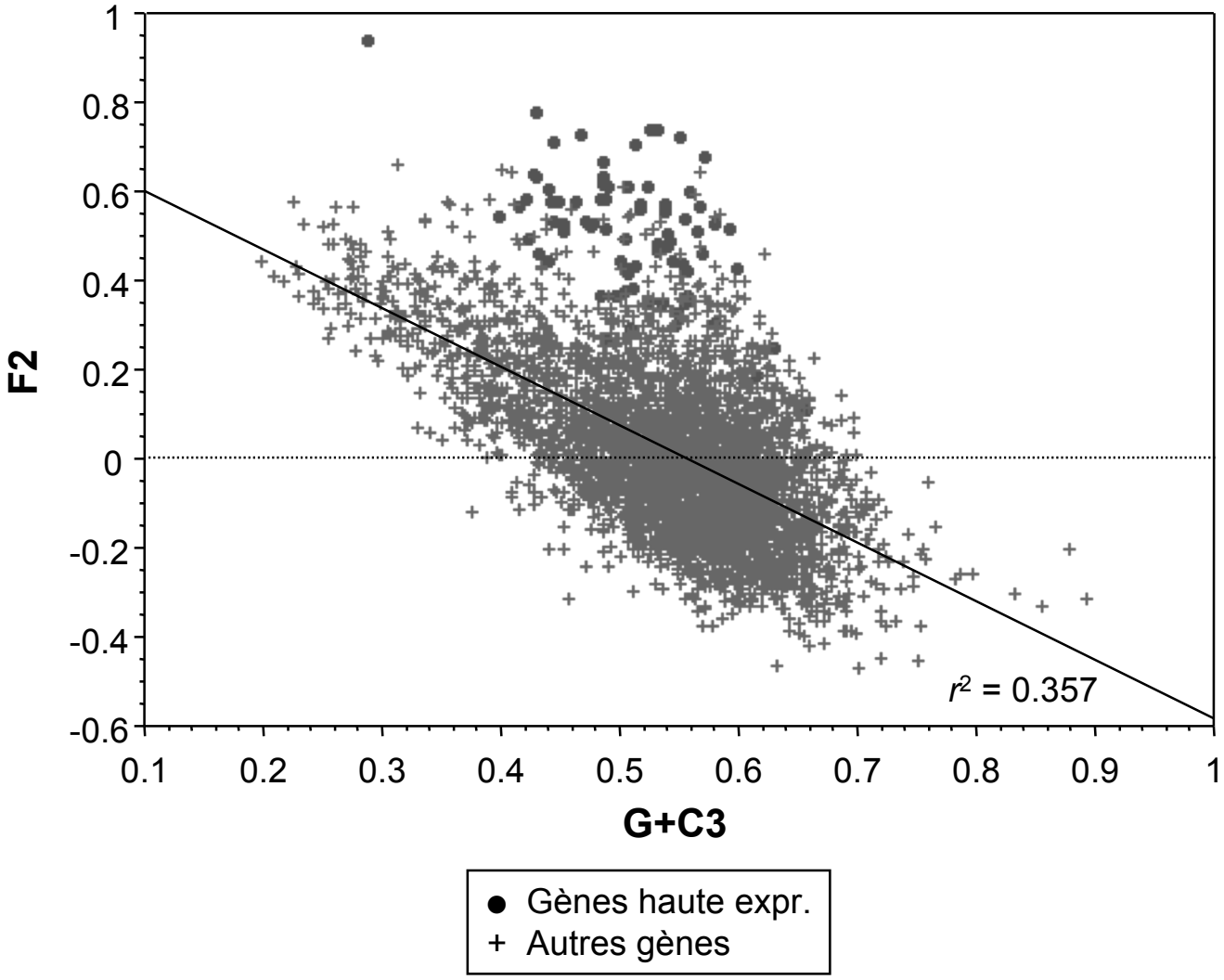
Résultats chez *E. coli*



Premier facteur de l'AFC



Deuxième facteur de l'AFC



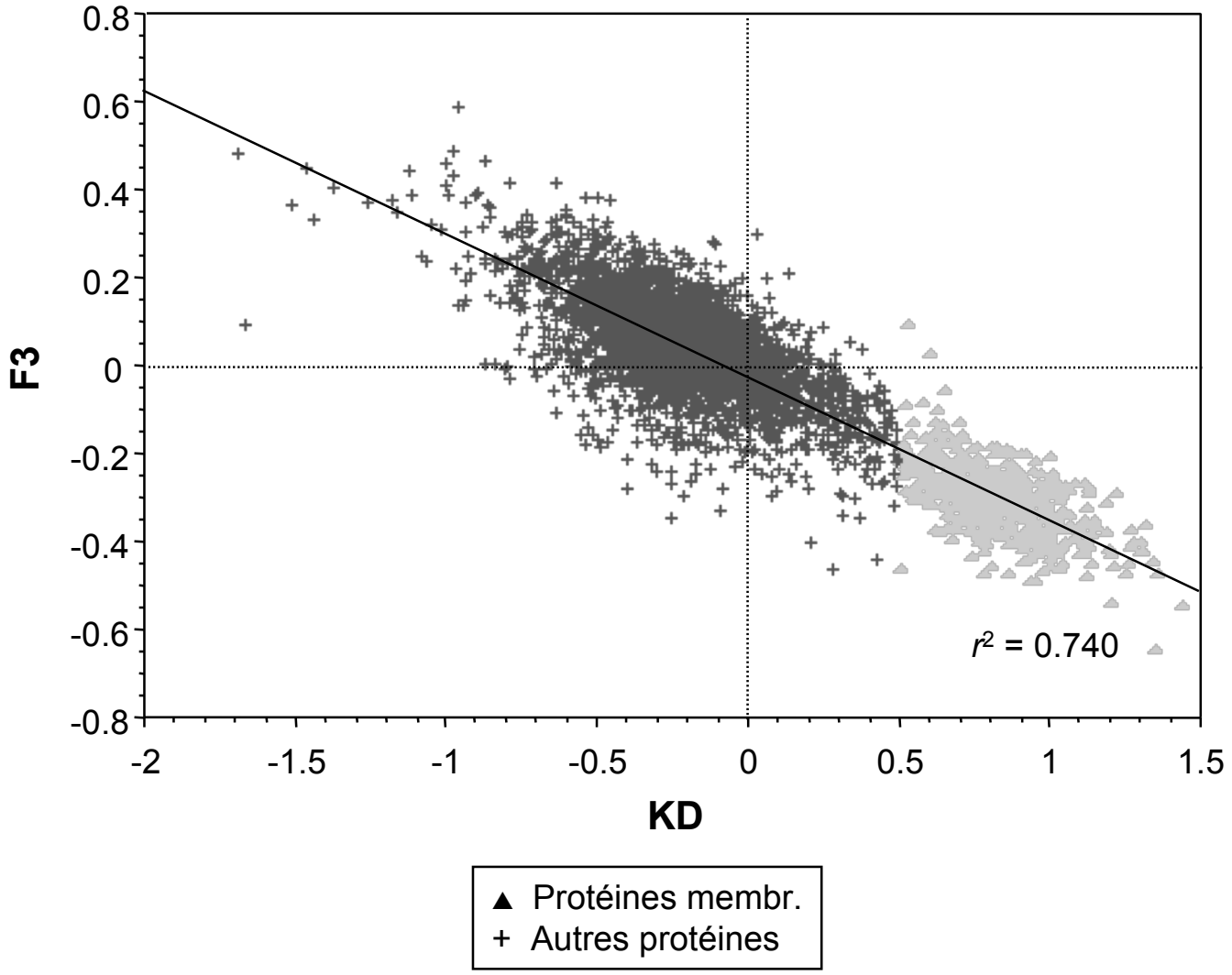
L'indice de Kyte & Doolittle

- La valeur de l'hydrophobicité d'une protéine (selon Kyte et Doolittle) est donnée par :

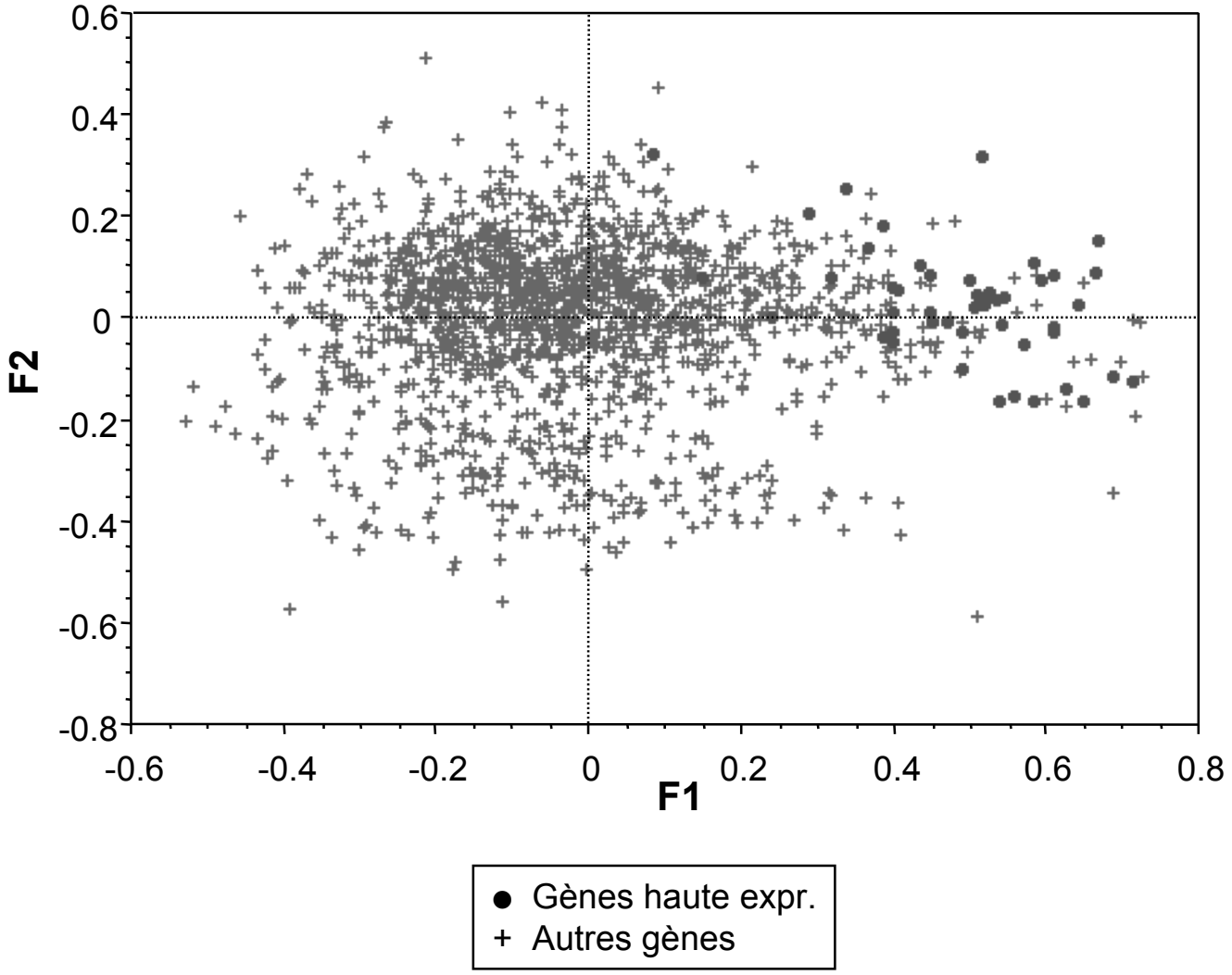
$$\mathrm{KD} = \frac{1}{n} \sum_{i=1}^{61} n_i \phi_i$$

où n est le nombre de codons du gène, n_i le nombre de codons i, et ϕ_i l'hydrophobicité de l'acide aminé correspondant à ce codon.

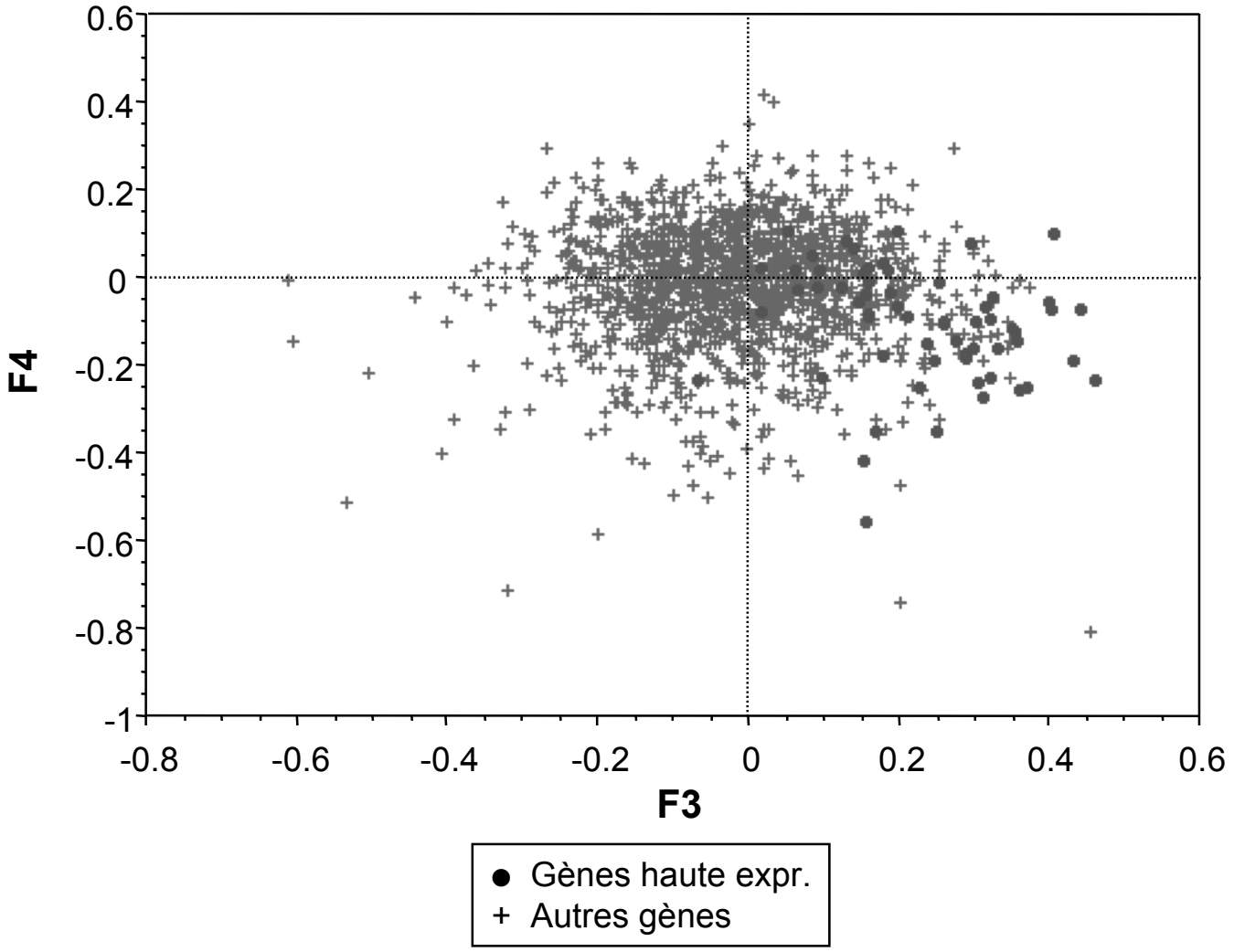
Troisième facteur de l'AFC



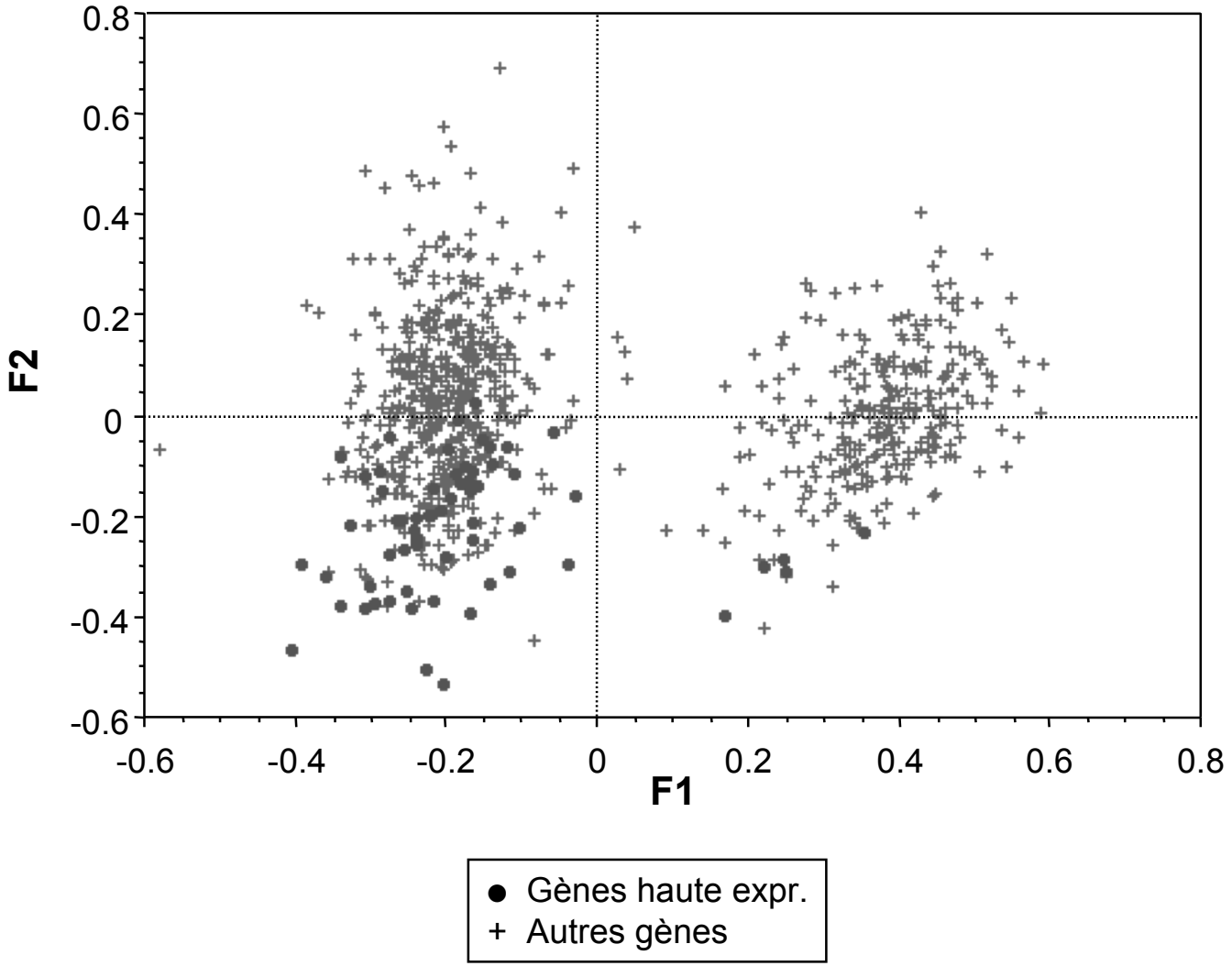
Résultats chez *H. influenzae*



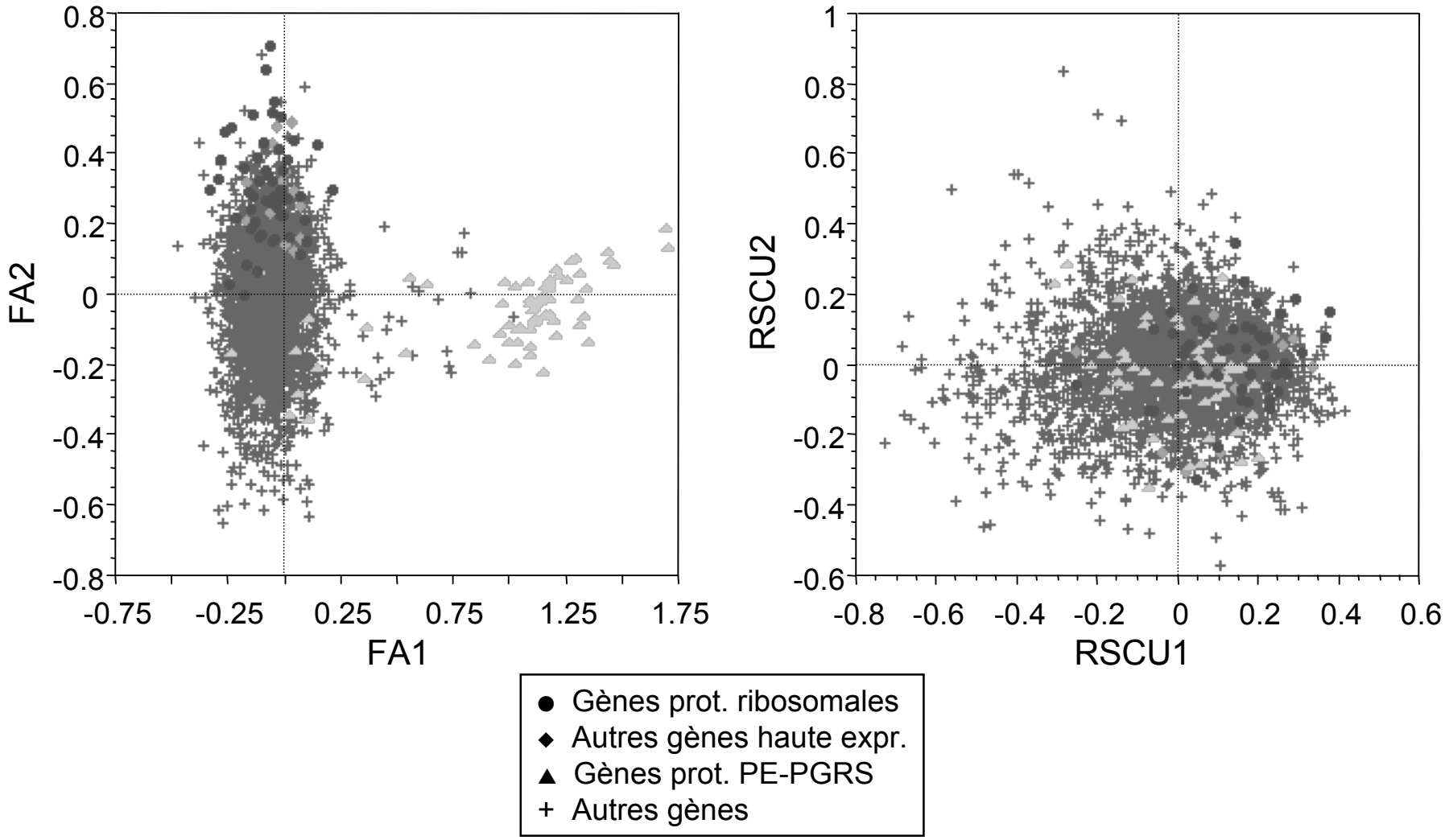
Résultats chez *H. pylori*



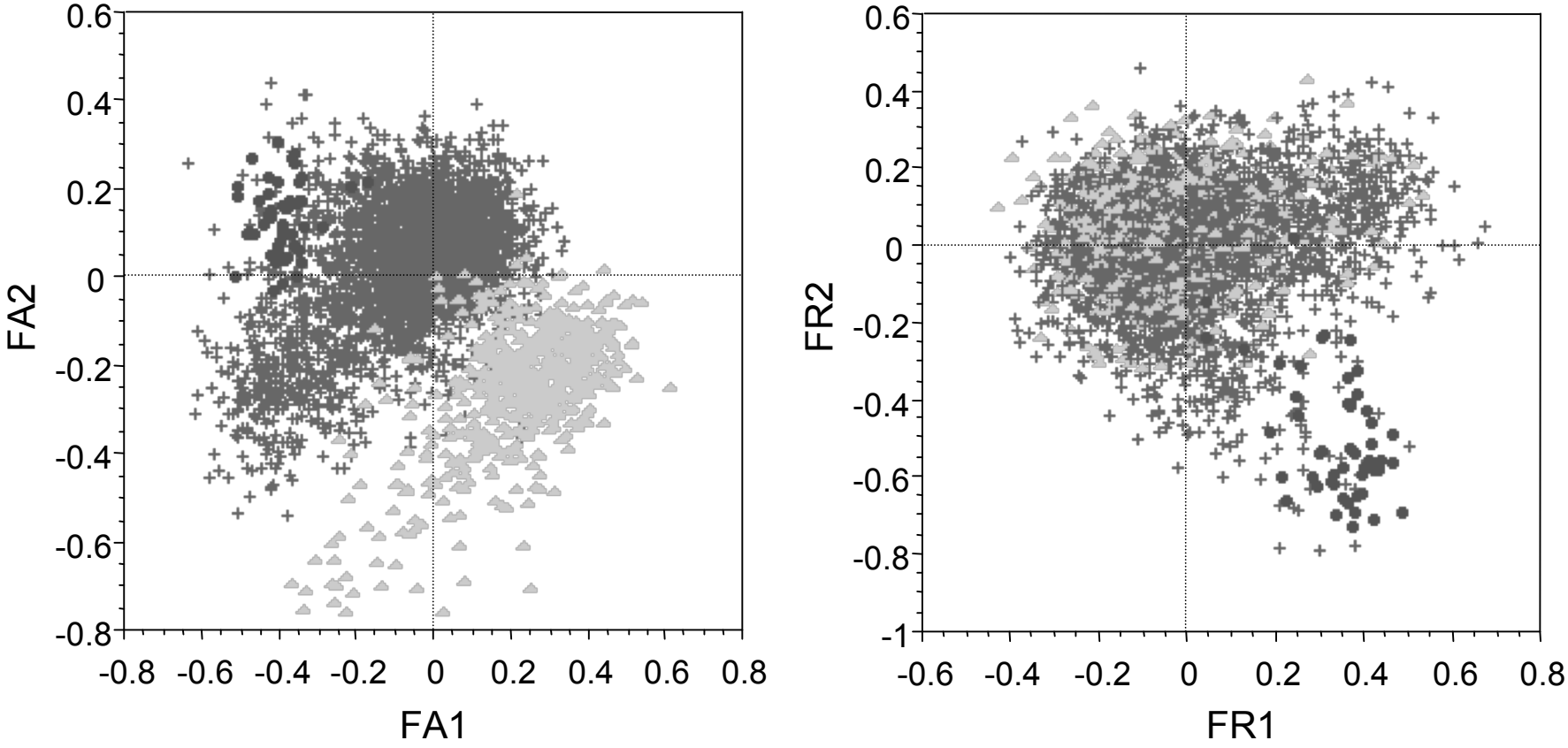
Résultats chez *B. burgdorferi*



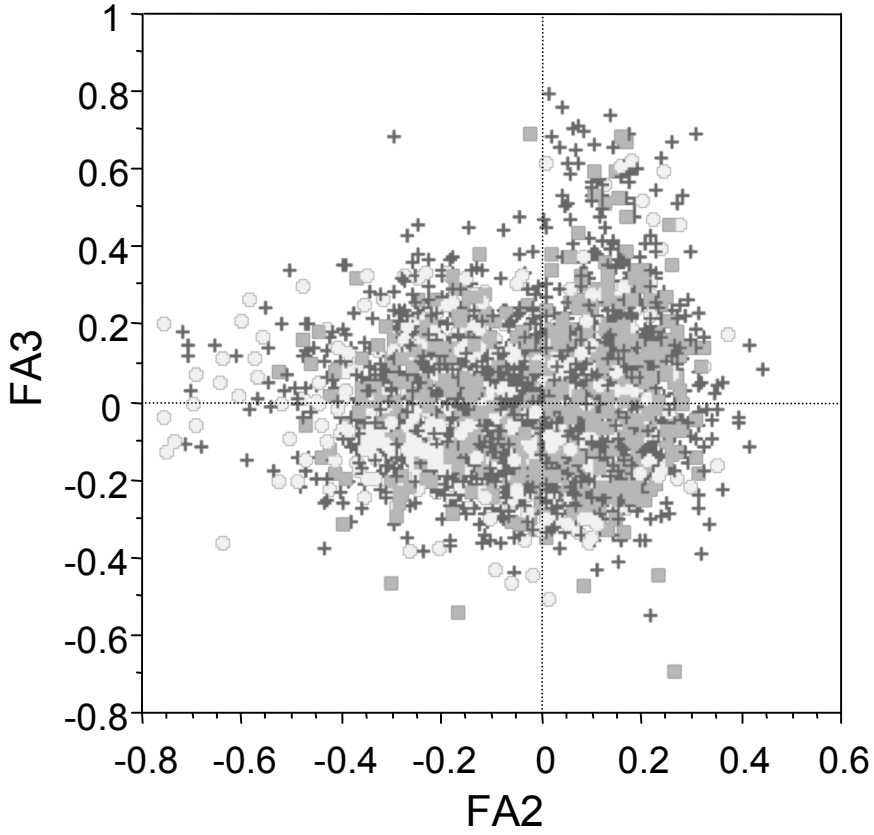
Résultats chez *M. tuberculosis*

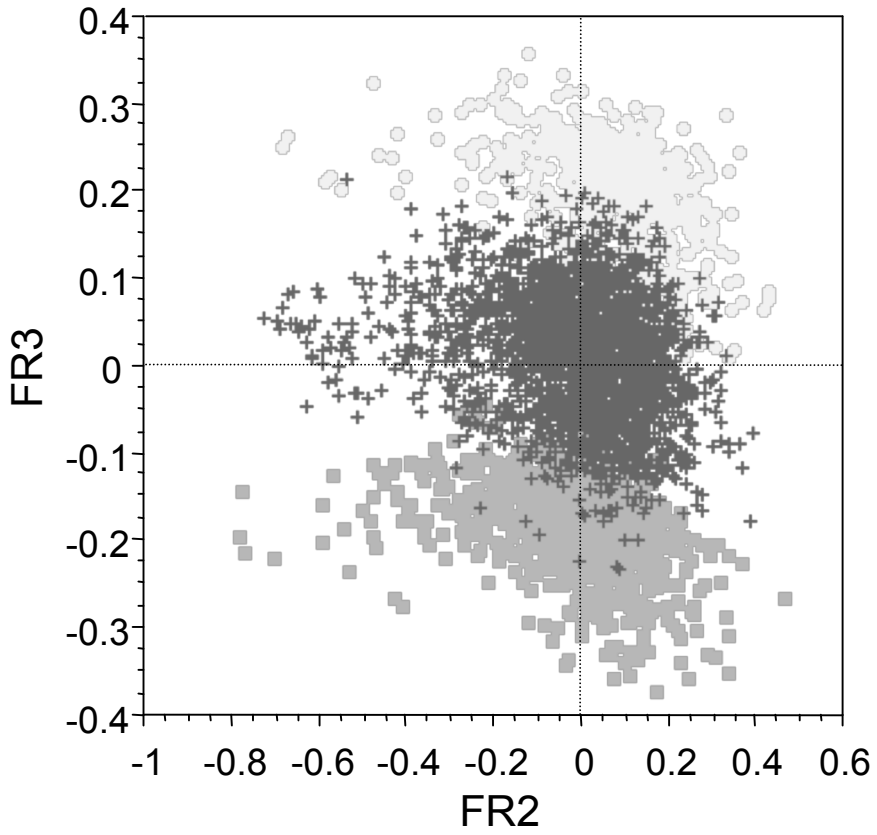


Résultats chez *B. subtilis*



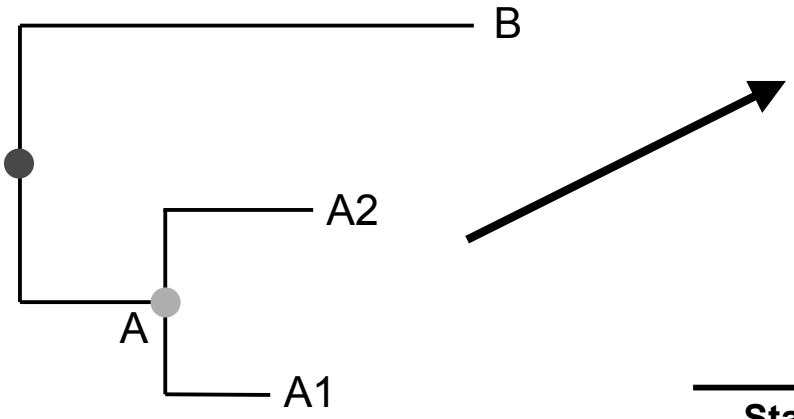
Résultats chez *B. subtilis*

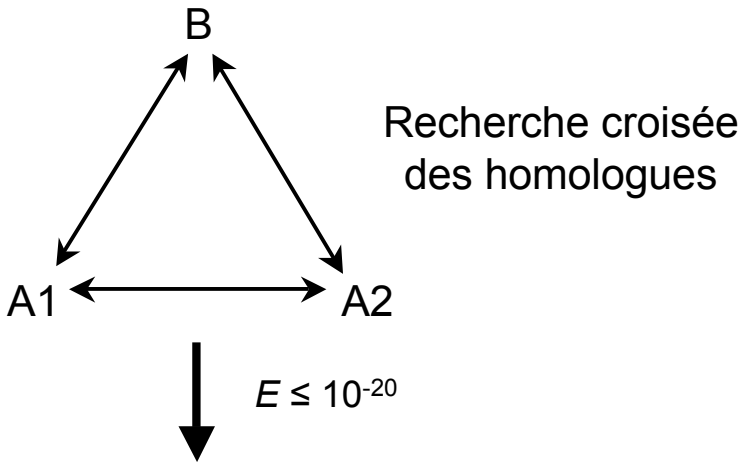




● Fréq. relatives UGC/UGU =1/0
■ Fréq. relatives UGC/UGU = 0/1
\+ Autres fréq. relatives

Approche comparative



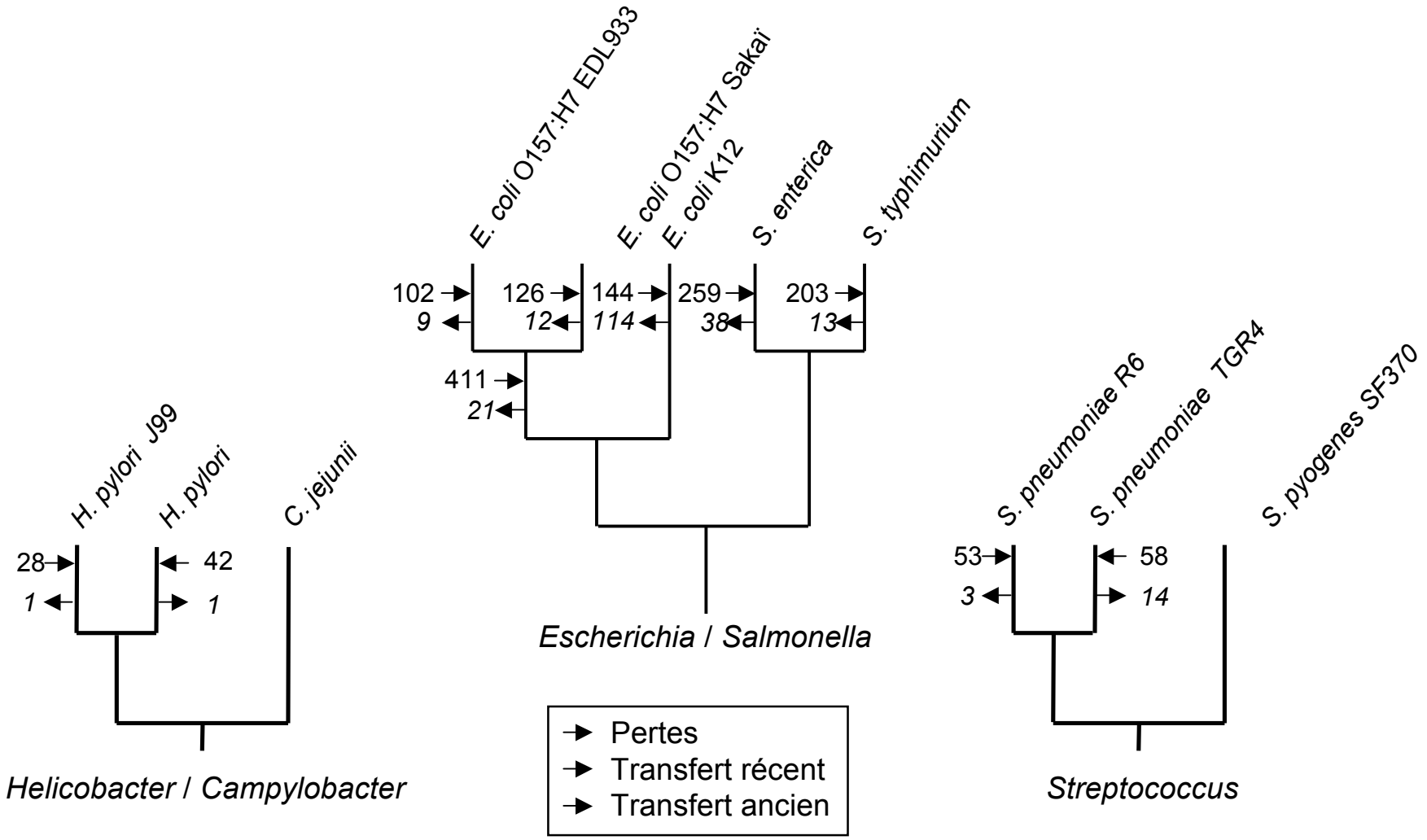


Divergence A/B (T_{AB})

Divergence A1/A2 (T_{12})

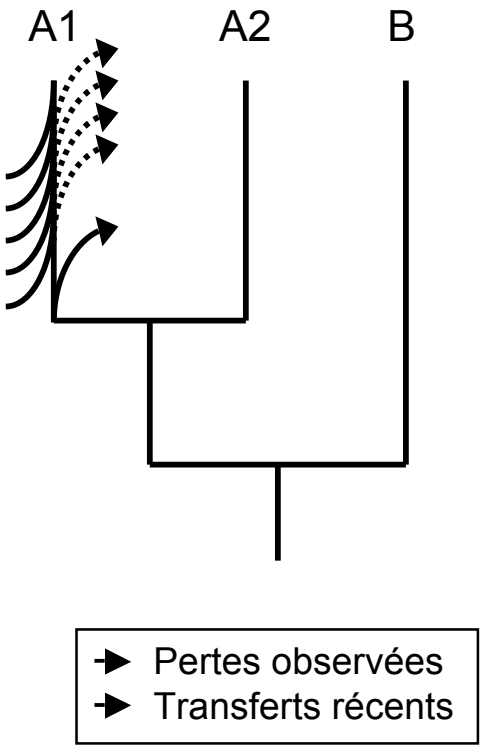
Statut	Événement probable
+A1 −A2 −B	Importation dans A1 après T_{12}
+A1 +A2 −B	Importation dans A après T_{AB} ou délétion dans B
+A1 −A2 +B	Délétion dans A2 après T_{12}
+A1 +A2 +B	Aucun événement
−A1 +A2 +B	Délétion dans A1 après T_{12}
−A1 +A2 −B	Importation dans A2 après T_{12}
−A1 −A2 +B	Délétion dans A après T_{AB}

Analyses effectuées



Gains et pertes

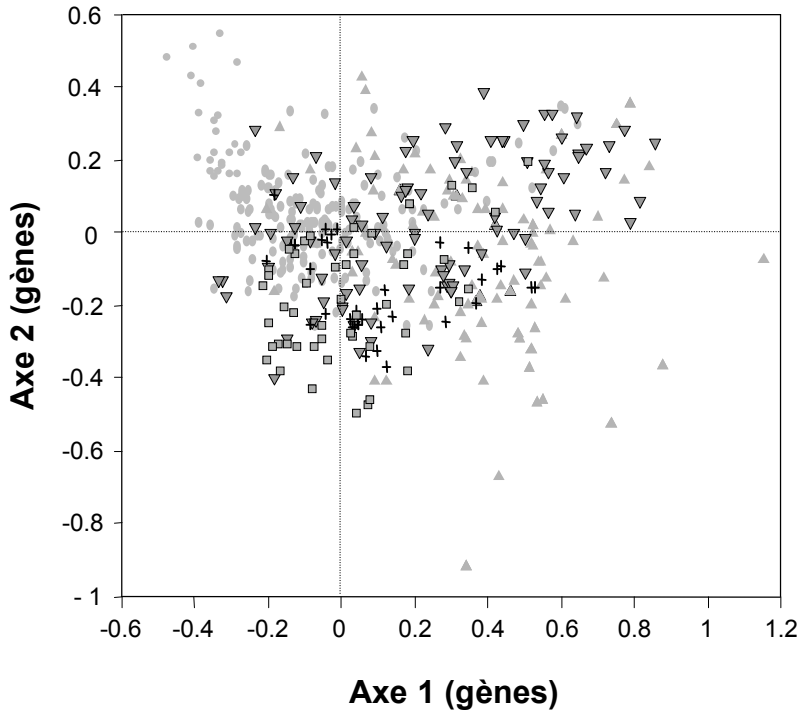
- Nb. d'acquisitions systématiquement supérieur au nb. de pertes :
  - Augmentation de la taille des génomes :
    - Cas des deux *E. coli* O157:H7 pathogènes.
  - Il existe un *turnover* rapide des gènes transférés:
    - Délétion peu de temps après transfert.

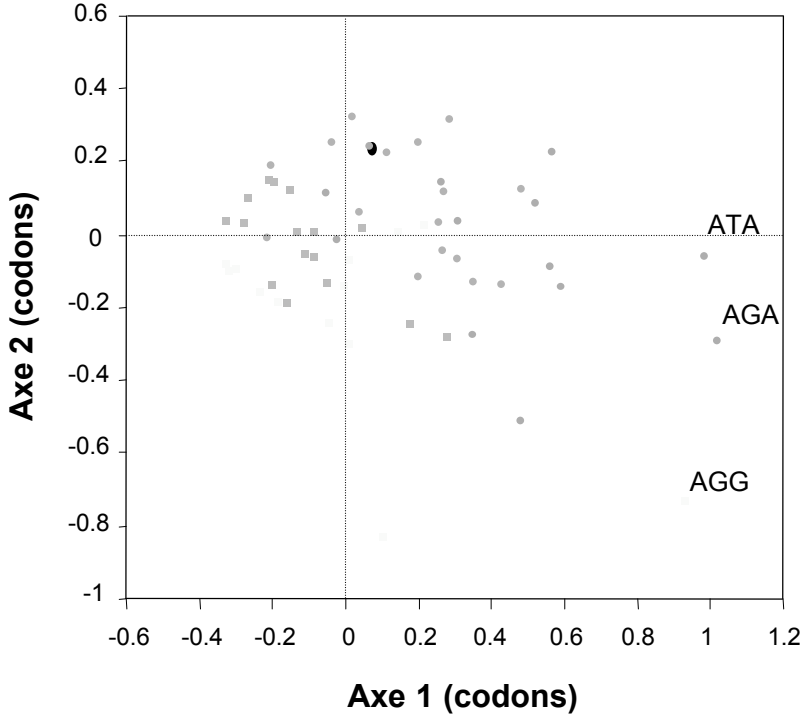


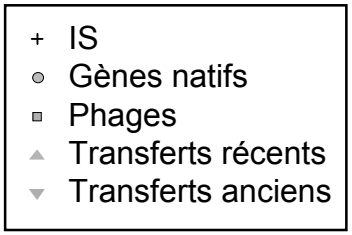
Étude de l'usage du code

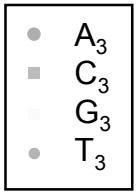
- Analyse par une AFC :
  - Séparation des gènes en cinq groupes :
    - Éléments IS, gènes natifs, phages et prophages, transferts récents et transferts anciens.
  - Tirage aléatoire de 200 représentants parmi l'ensemble des gènes appartenant à la classe « native ».
  - Tirage aléatoire de 50 représentants pour les phages et les éléments IS.
  - Dix répétitions effectuées pour vérifier la conservation des résultats avec des échantillons différents.

Escherichia coli

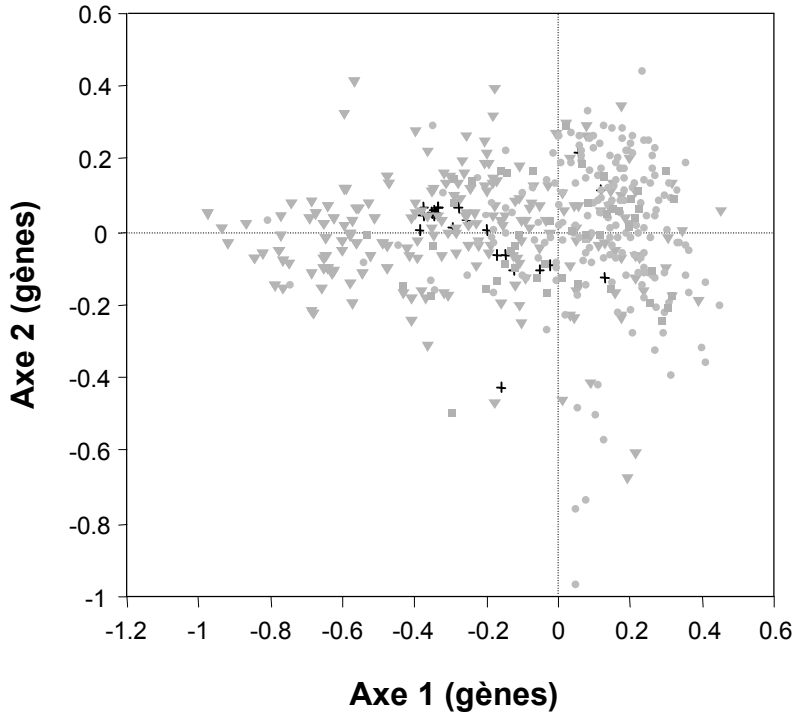


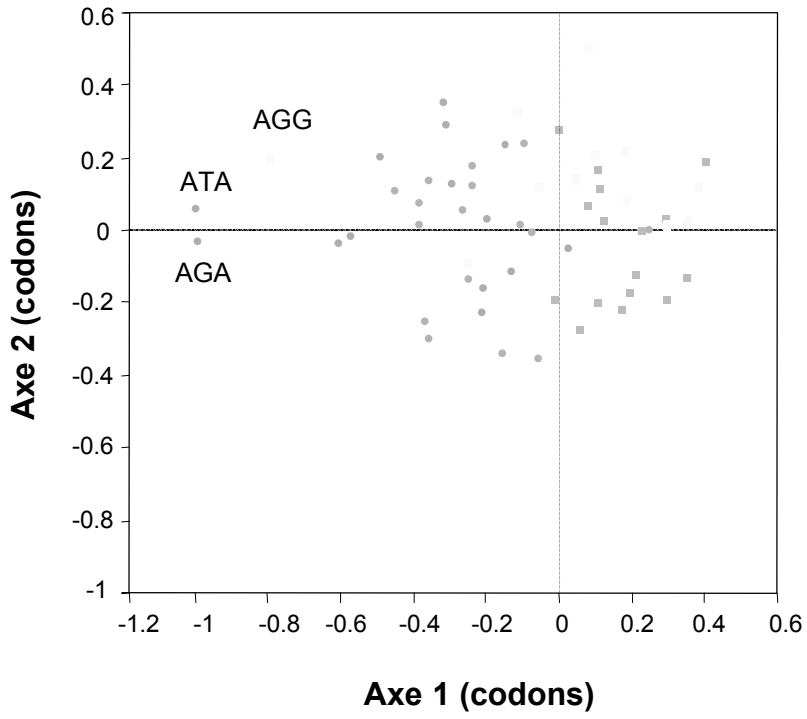


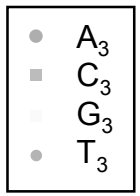




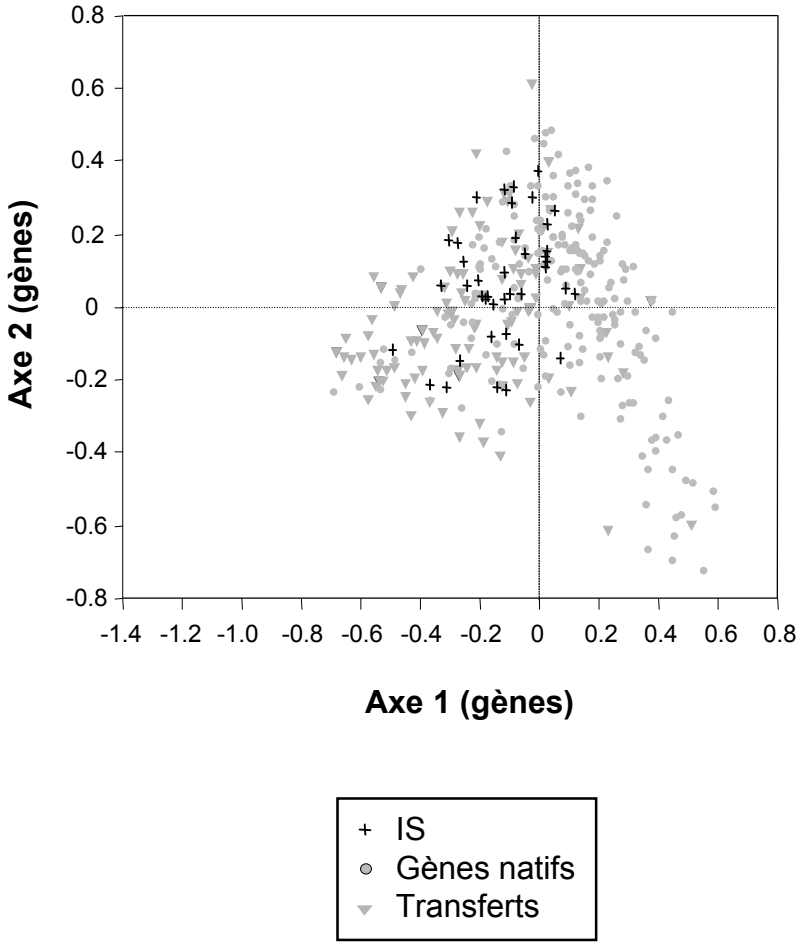
Salmonella enterica

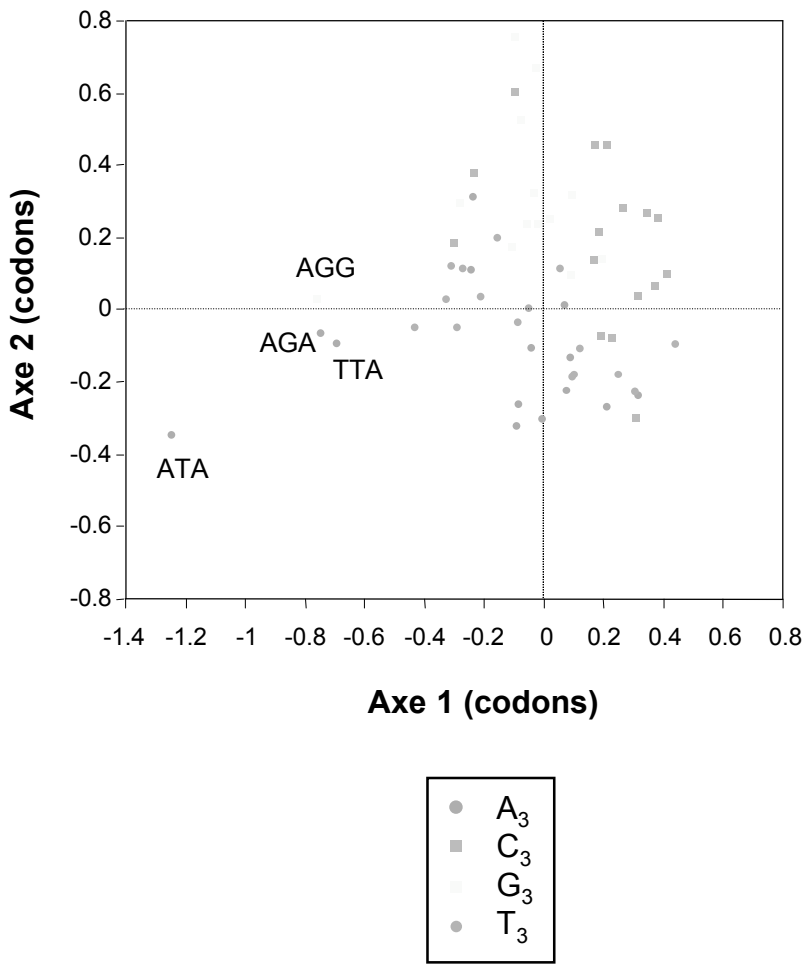




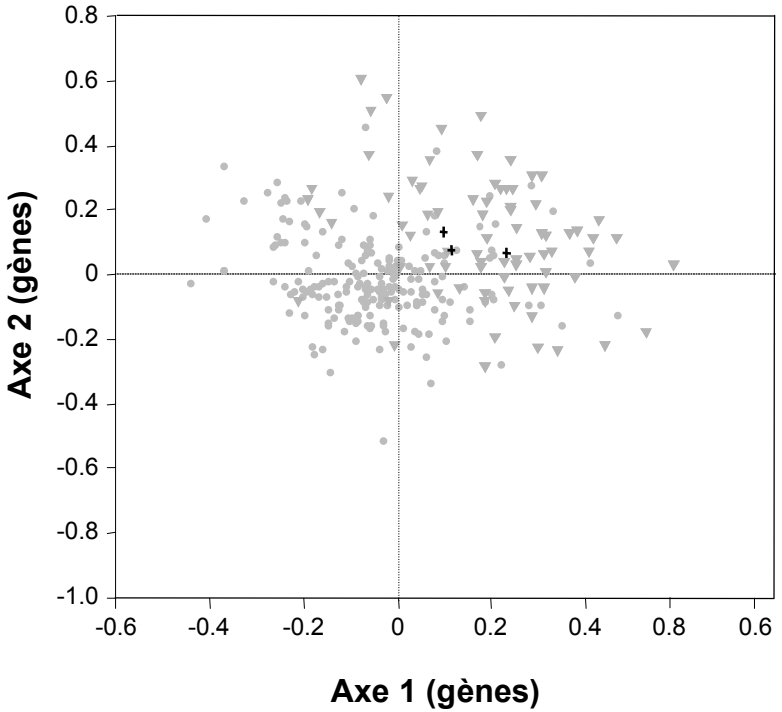


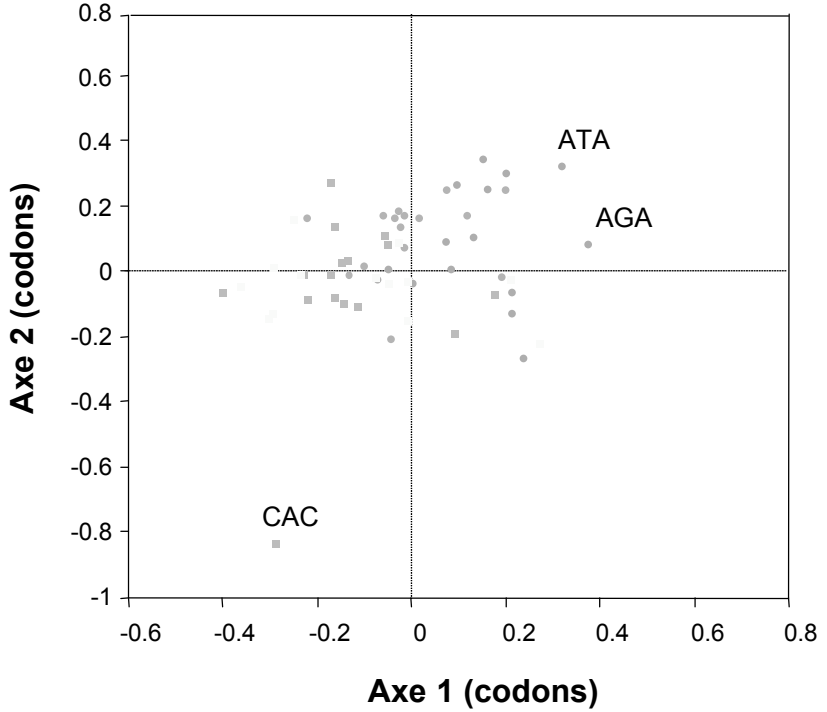
Streptococcus pneumoniae

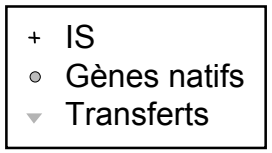


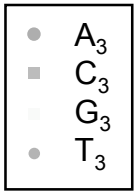


Helicobacter pylori

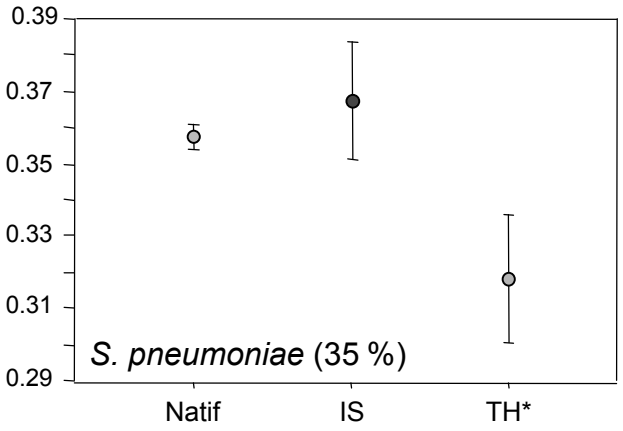


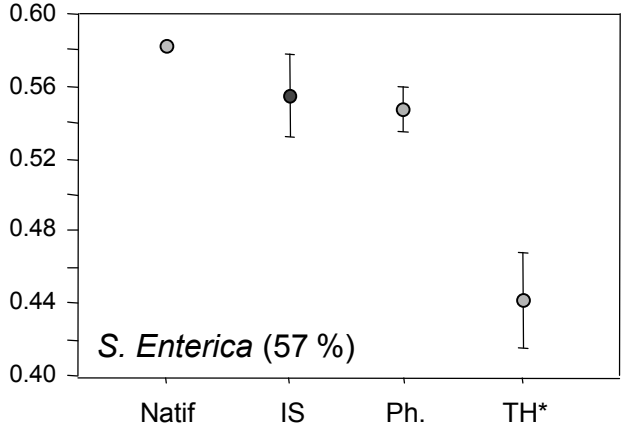


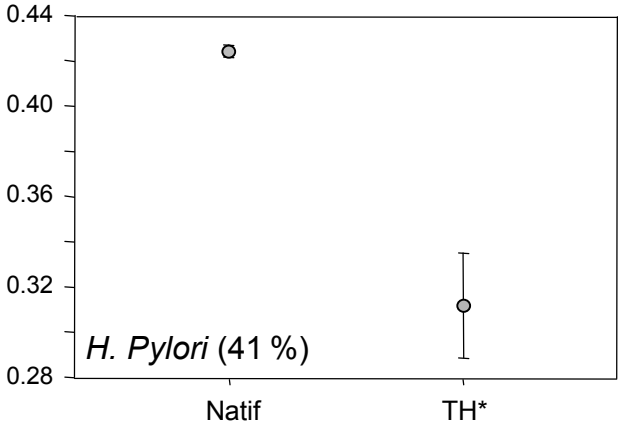


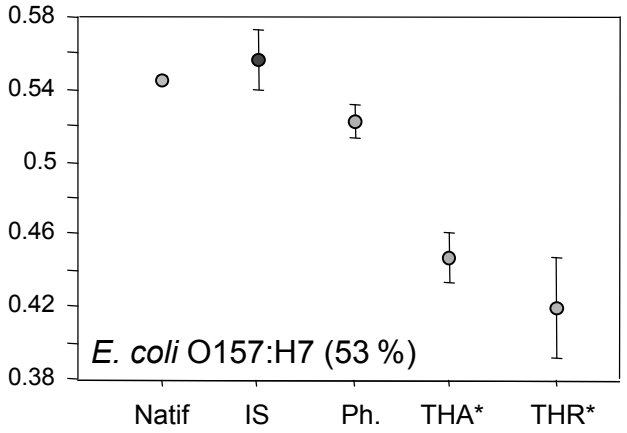


Contenu en G+C3









* Test de Mann-Whitney significatif à $P < 10^{-4}$

Hypothèses possibles

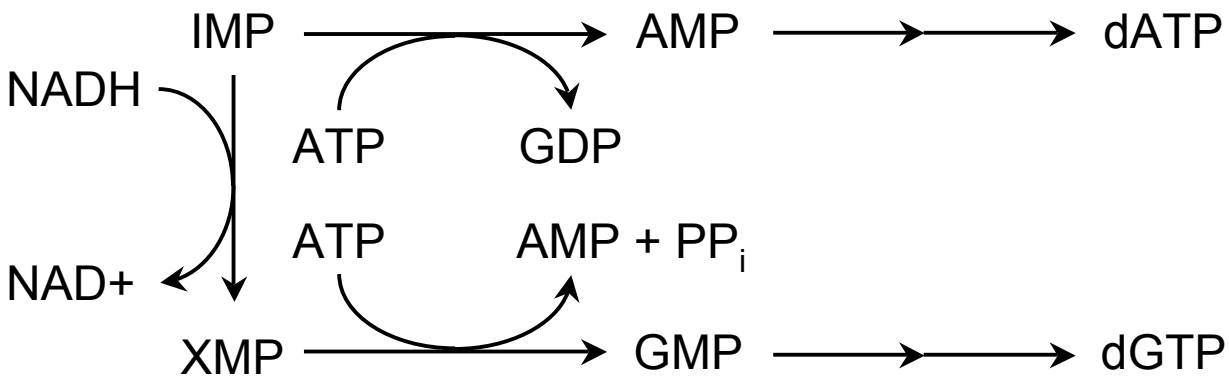
- Biais dans les séquences reconnues par les systèmes de restriction :
  - Celles-ci apparaissent comme étant globalement riches en G+C (70 % dans REBASE).
- Biais au niveau des vecteurs impliqués dans les transferts (phages et éléments IS) ?
- Les transferts horizontaux, un sous-produit des besoins en Adénine des bactéries ?

Compétition métabolique

- Parmi les ribonucléosides triphosphates (NTP), l'ATP est le plus abondant :
  - Au cœur de toutes les voies métaboliques.
- De tous les désoxyribonucléosides triphosphates (dNTP), le dCTP est le plus coûteux à synthétiser.
- La transformation, un mécanisme biaisé vers la capture d'ADN riche en Adénine ?

Biosynthèse des dNTP

Purines :



Pyrimidines :