Phylogénomique : introduction

Les limites des phylogénies basées sur un seul gène

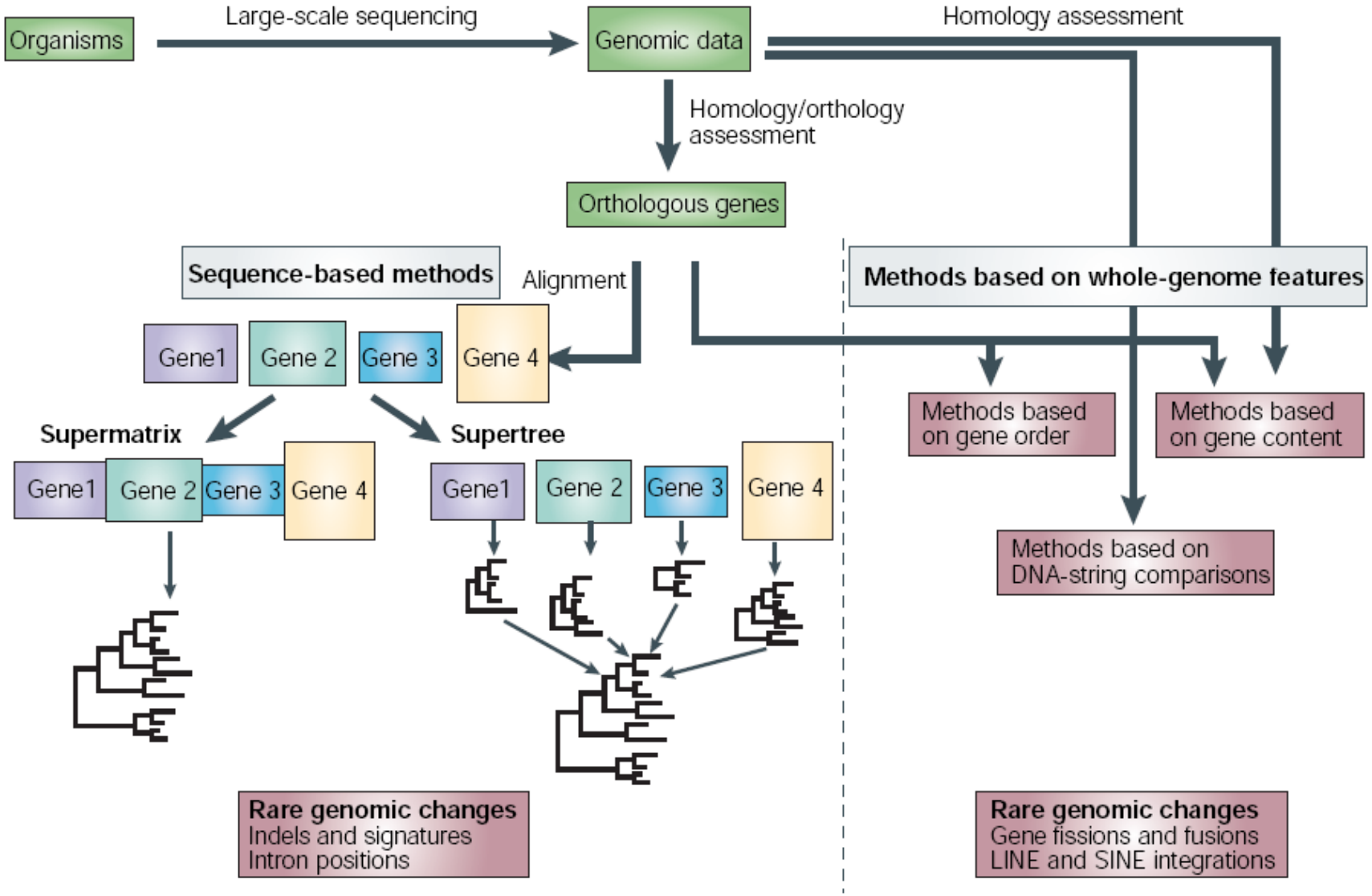
- **Résolution limitée (erreur stochastique)**

- **La phylogénie du gène peut être différente de la phylogénie des espèces à cause de :**
  - **La paralogie cachée**
  - **Des transferts horizontaux de gènes**
  - **Du polymorphisme ancestral**

➡ **Inférer les phylogénies à partir de caractères génomiques**

Les différentes méthodes d'analyse phylogénomique

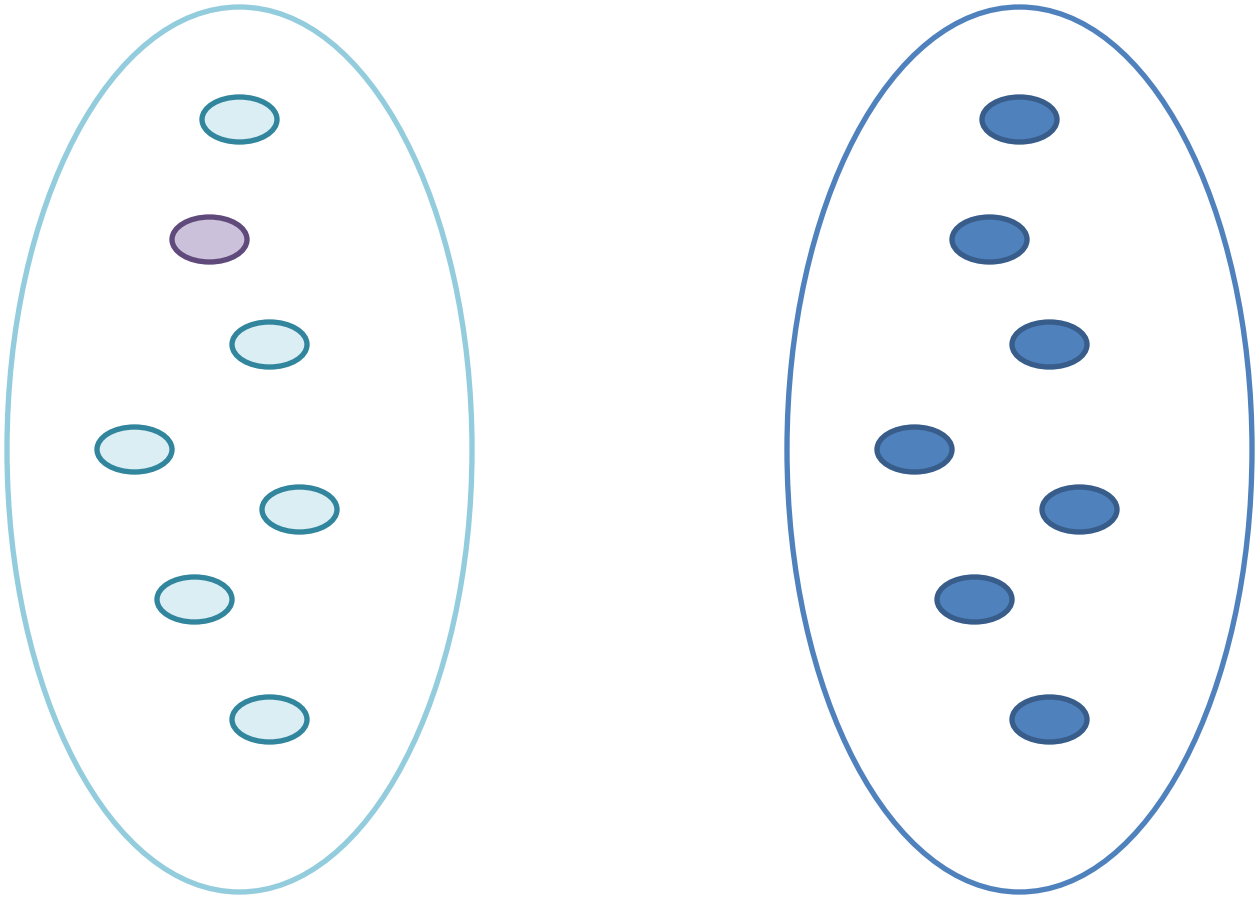
(Extrait de Delsuc *et al.*, 2005, *Nat. Rev. Genet. 6: 361-375)*



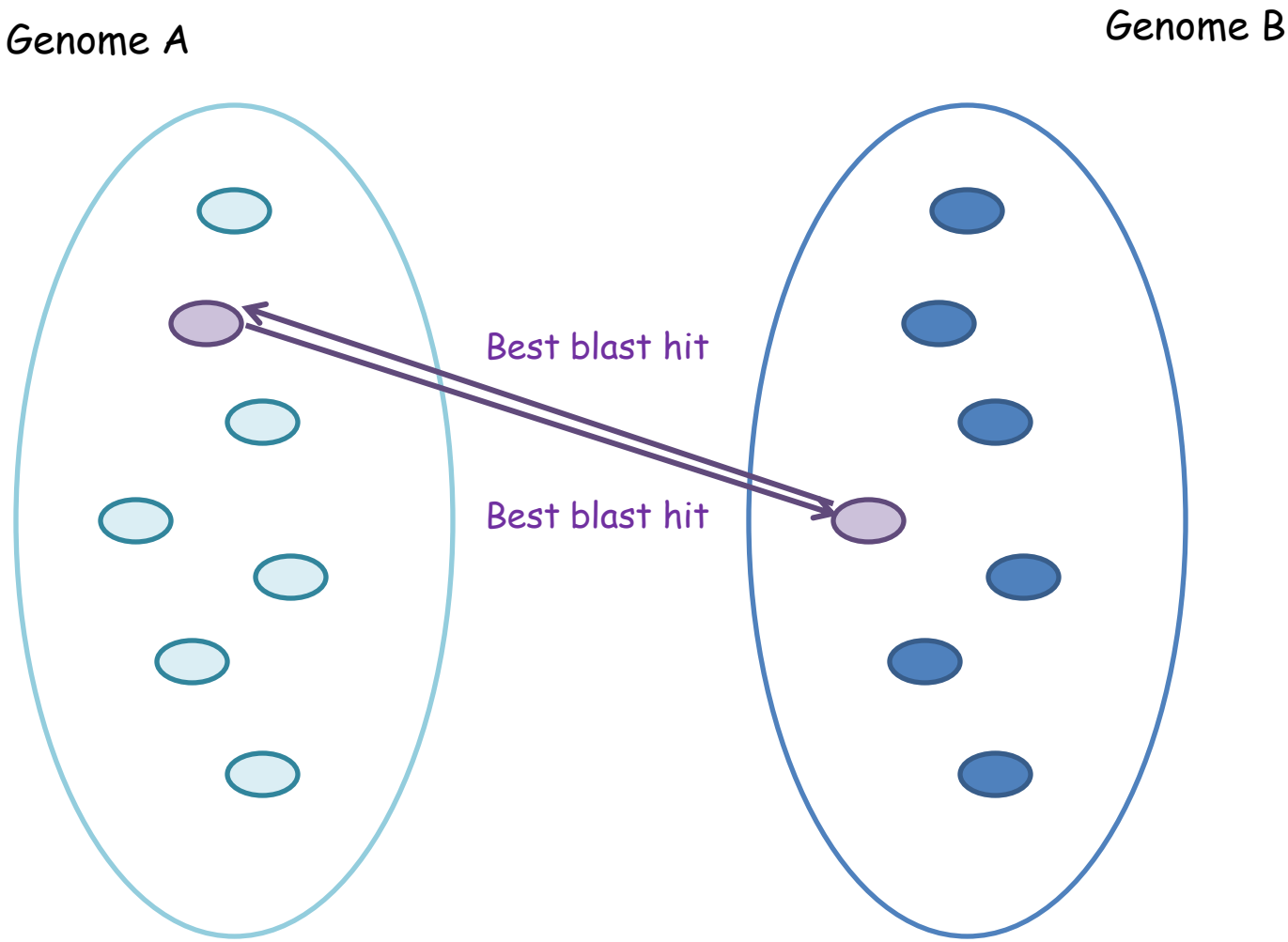
Orthologie en pratique

Genome A

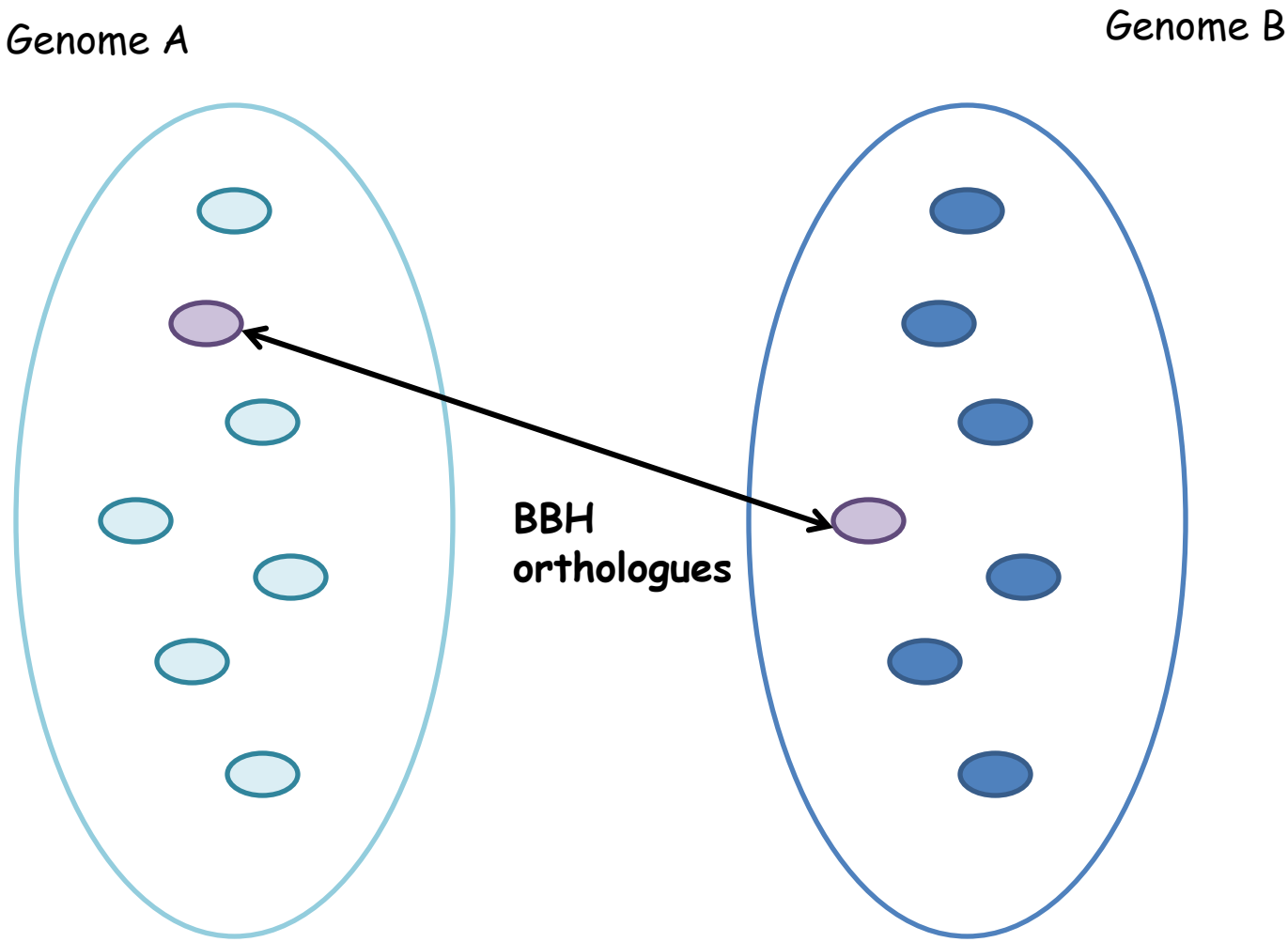
Genome B



Orthologie en pratique



Orthologie en pratique



Méthodes basées sur le contenu en gènes

- recherche des gènes orthologues entre les génomes analysés

- La notion de similarité entre deux espèces est définie comme le rapport du nombre de gènes orthologues communs sur une valeur normalisée qui reflète la différence des tailles des génomes

Définition de la taille du génome :

- nombre d'ORFs annotés
- nombre d'ORFs avec au moins un homologue dans les autres génomes complets. Elimine la variabilité des prédictions de gènes. Meilleur estimateur
- nombre d'ORFs avec au moins un orthologue dans les autres génomes complets. Option stringente qui peut poser problème pour les génomes ayant subis des duplications de gènes

Normalisation :

- Division par la taille du plus petit génome (maximum théorique d'orthologues partagés). Cette mesure a tendance à regrouper dans un même cluster les petits génomes lors de la construction de l'arbre.
- Division par la moyenne pondérée des tailles dont la valeur est : $\frac{ab\sqrt{2}}{\sqrt{a^2+b^2}}$

Méthodes basées sur le contenu en gènes

Mesure de la distance évolutive entre génomes :

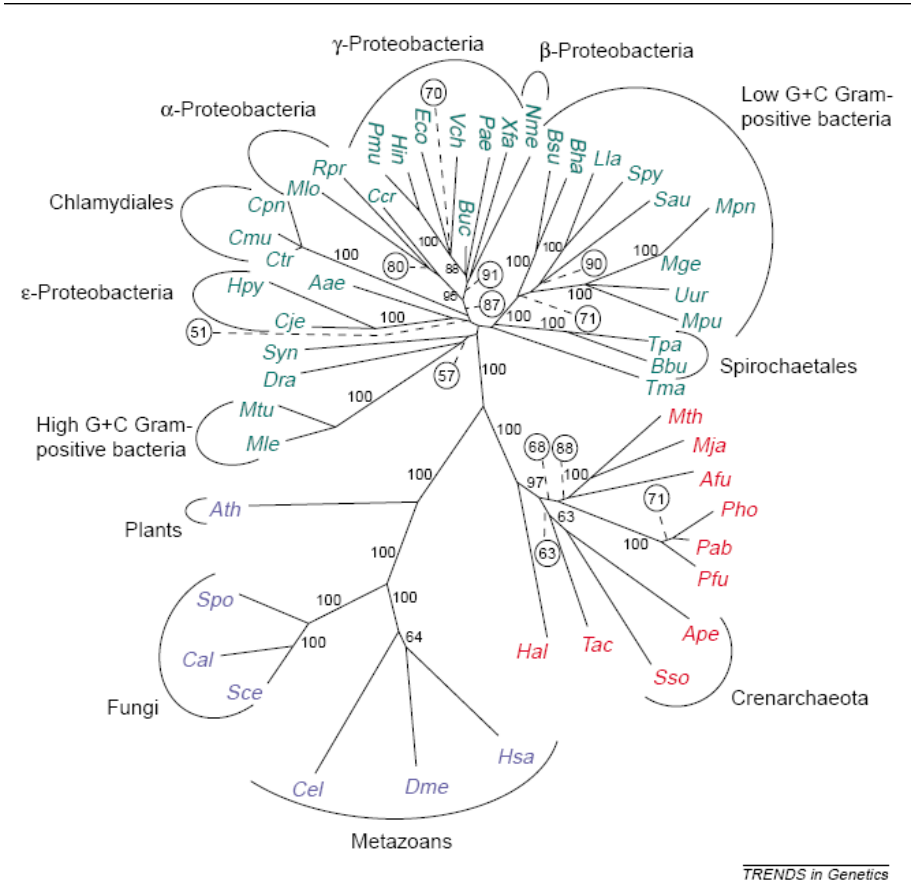
Basée sur la similarité estimé *s* → $d = -ln(s)$

Construction de l'arbre :

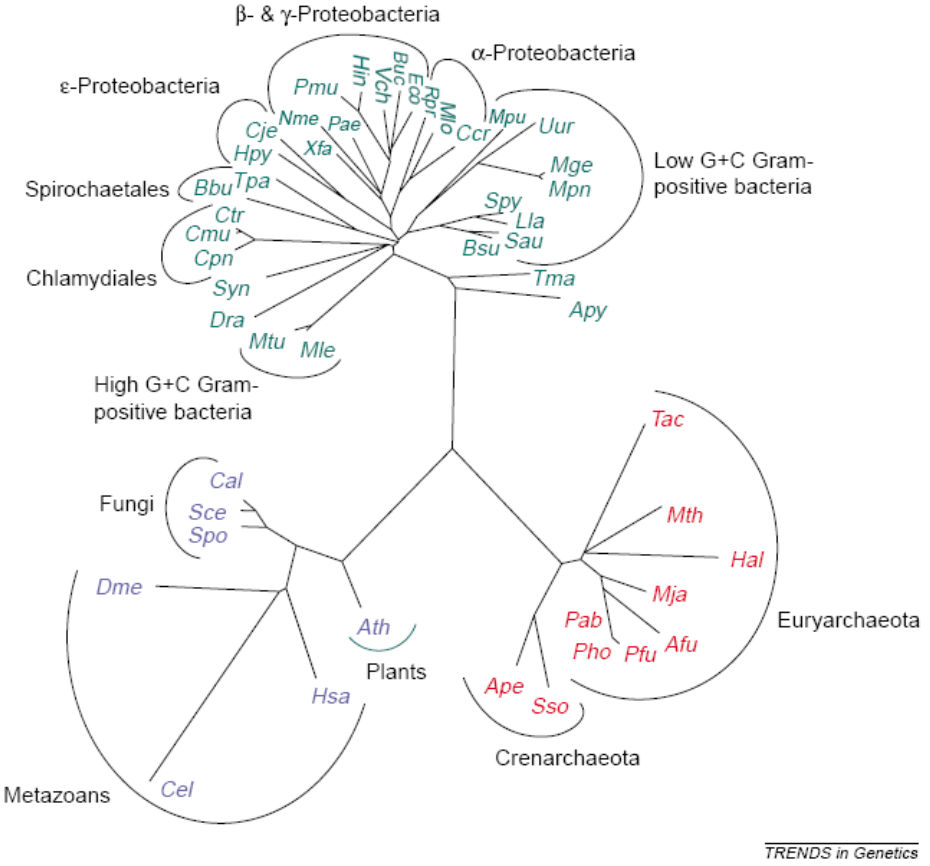
Méthode de distances : La Neighbour-joining

Méthodes basées sur le contenu en gènes

(Extrait de Korbel et al. (2002) Trends Genet. 18 : 158-162)



Arbre obtenu sur le nombre de gènes communs entre 50 génomes complets



Arbre obtenu sur l'ANRr de la petite sous unité ribosomique

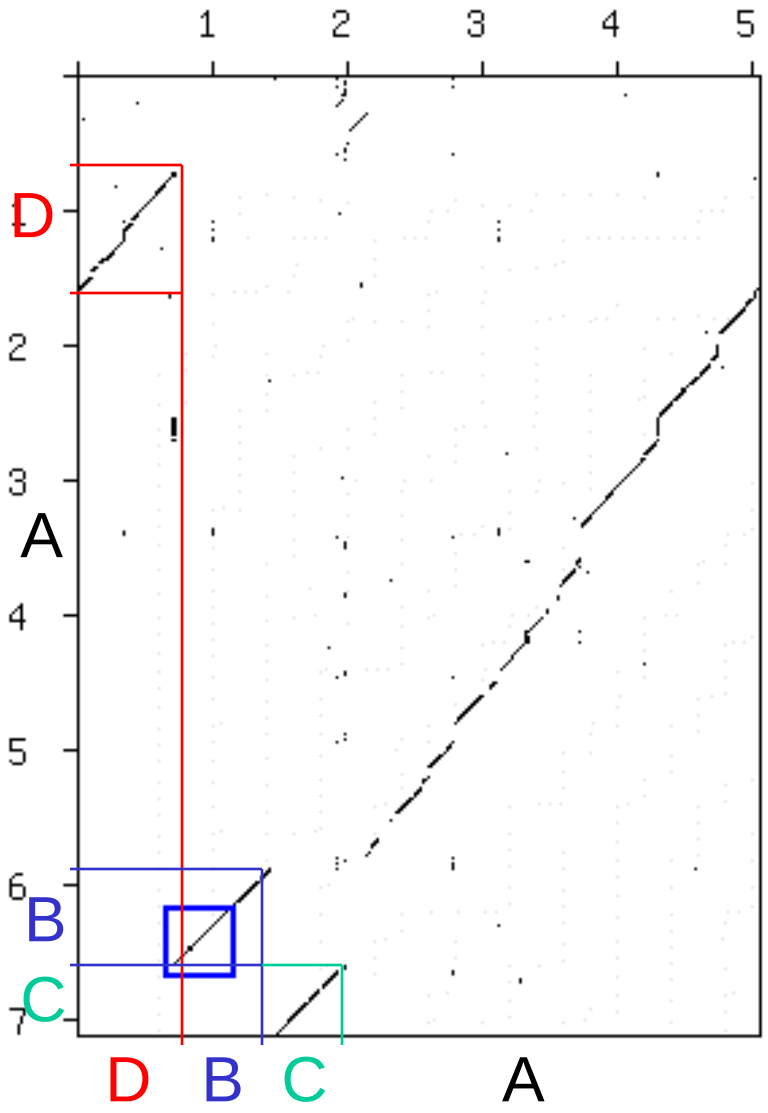
Méthodes basées sur la conservation de l'ordre des gènes

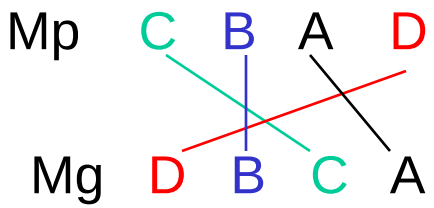
Utilisées essentiellement pour les génomes procaryotes

Pourquoi cette mesure?

Mycoplasma genitalium

Mycoplasma pneumoniae

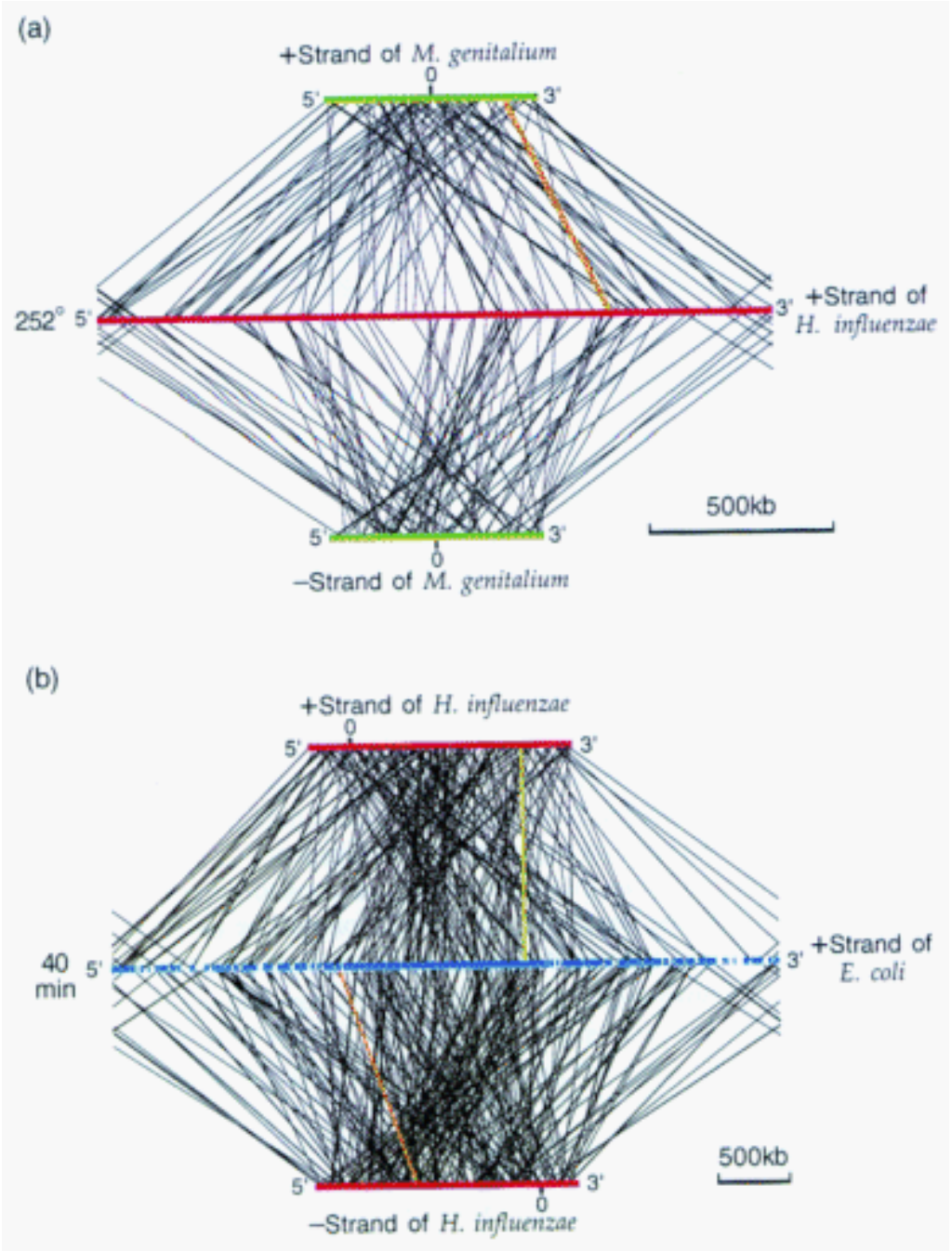




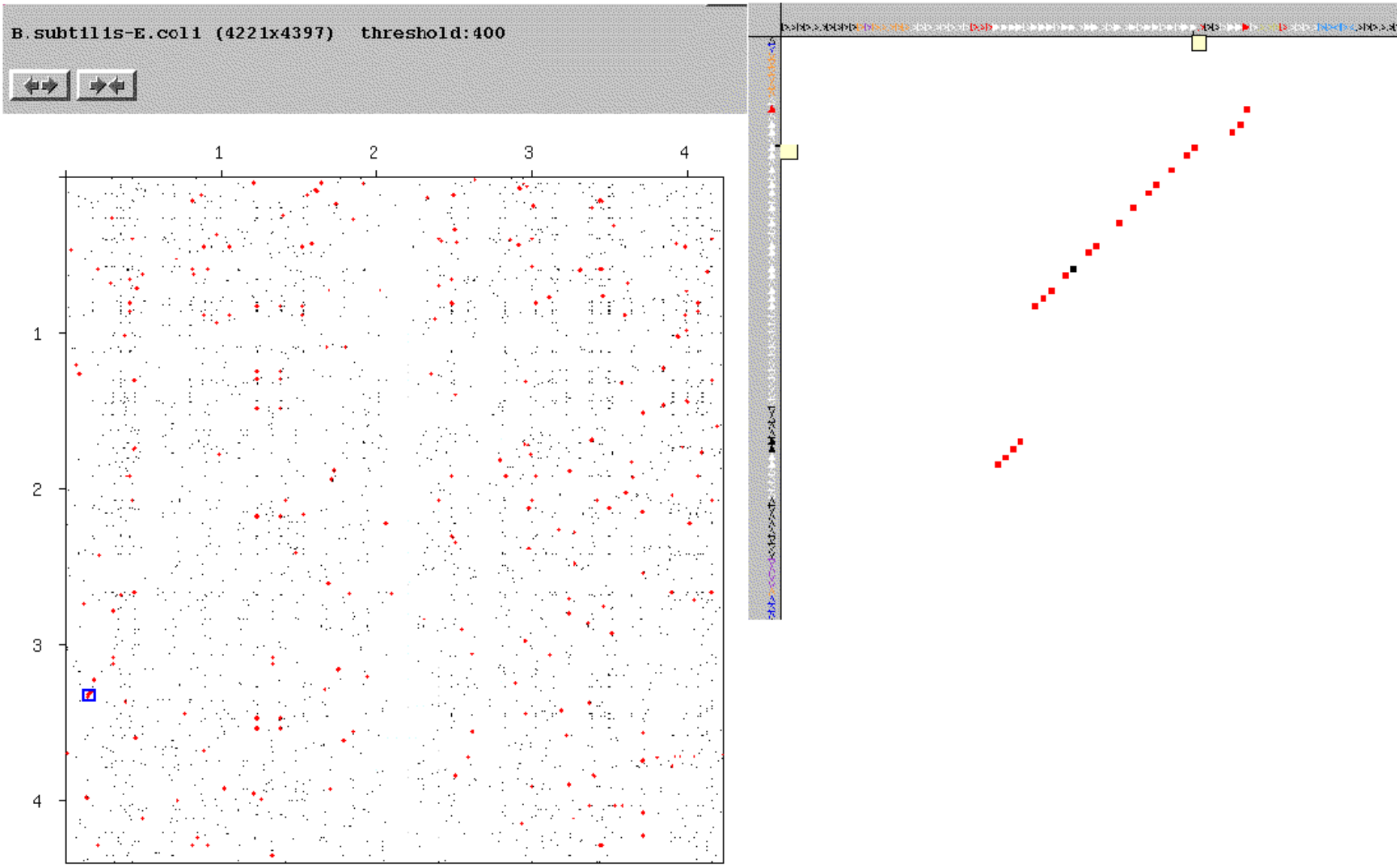
Comparaison de la position des gènes orthologues entre deux génomes

a) *M. genitalium/H. influenzae*
b) *H. influenzae/E. coli*

(extrait de Watanabe et al., 1997,
J. Mol. Evol., 44 (Suppl. 1, S57-S64)

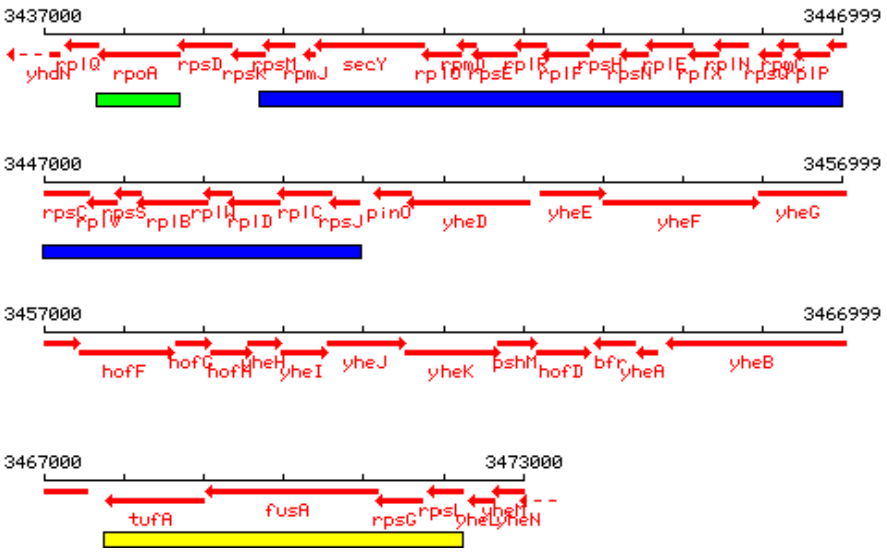


Conservation des gènes orthologues entre *E. coli* et *B. subtilis* (issus de KEGG)

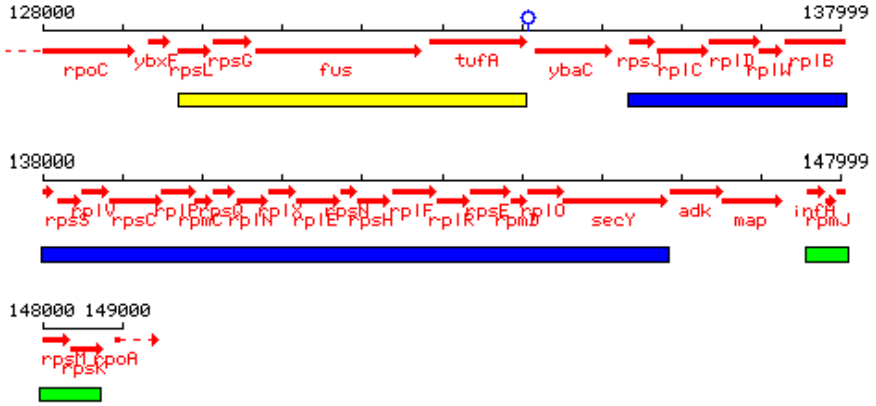


Comparaison de l'organisation des gènes de l'opéron des protéines ribosomiques (issus de KEGG)

Escherichia coli



Bacilus subtilis



Conservation de la structure en opérons au sein des génomes

(Itoh et al. (1999) Mol. Biol. Evol., vol 16, 332-346)

Analyse réalisée à partir de structures opéroniques déterminées expérimentalement :

- 256 opérons décrits chez *Escherichia coli* (~ 3,5 gènes/opéron)
- 100 opérons décrits chez *Bacillus subtilis* (~4,1 gènes/opéron)

La comparaison de ces opérons avec les opérons orthologues de 11 autres génomes complètement séquencés a été réalisée. Le génome de *Saccharomyces cerevisiae* a été inclus dans cette liste car l'organisation de certains de ces gènes est connue pour être similaire à celle observée chez les bactéries.

Conservation de la structure en opérons au sein des génomes (Itoh et al., 1999)

Définition des paires d'orthologues:

- Les ORFS entre les deux génomes comparés doivent être les plus similaires de façon réciproque (BBH) (Fig. 1a).
- La similarité doit être statistiquement significative.
- Si un orthologue donné présente plus de similarité avec des paralogues du même génome, tous les paralogues sont considérés comme étant orthologues au partenaire de l'autre génome (Fig. 1b).

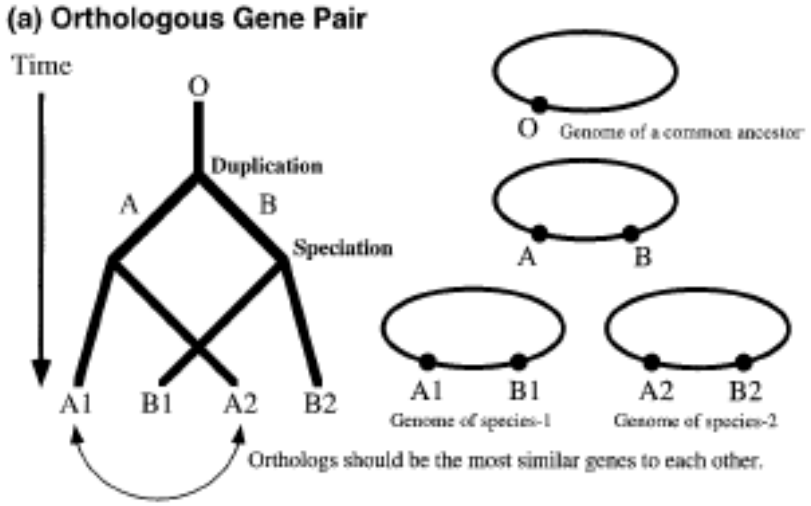


Fig. 1a

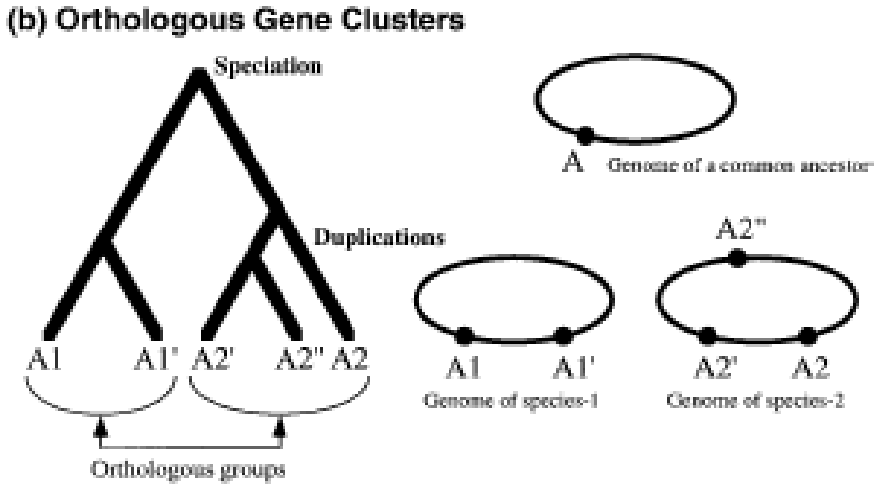
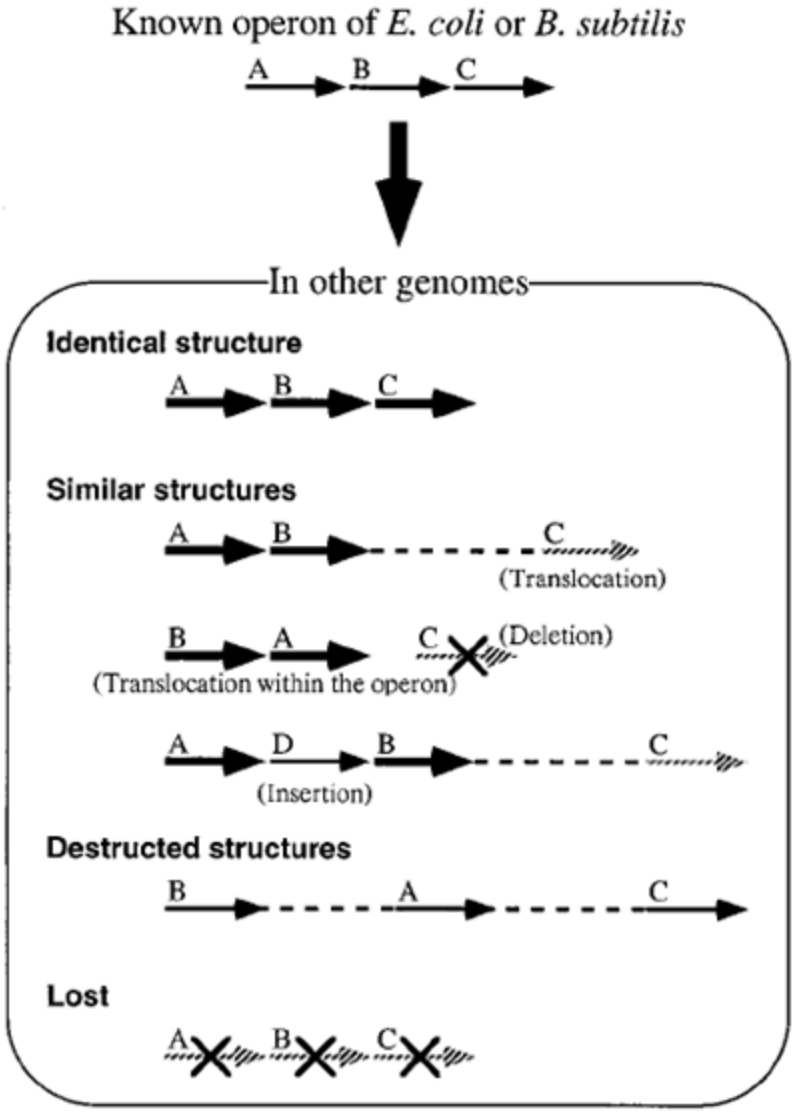


Fig. 1b

Conservation de la structure en opérons au sein des génomes (Itoh et al., 1999)



Définition des classes d'opérons orthologues

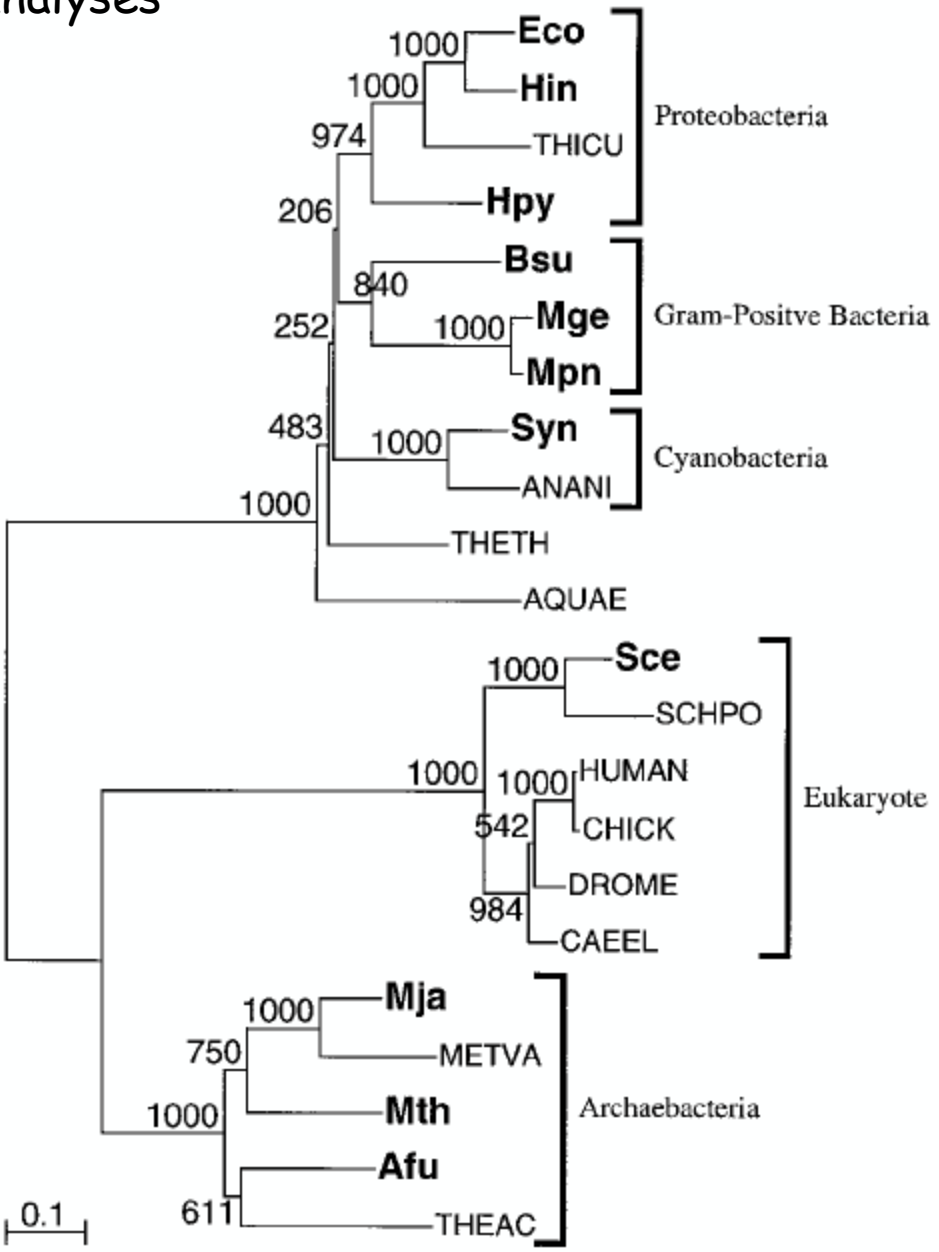
- structure opéronique identique (exactement la même dans les 2 génomes).
- structure opéronique similaire: partiellement conservée, les translocations, les délétions et au plus 2 insertions dans l'opéron sont autorisées.
- structure opéronique détruite: si 2 orthologues ou plus sont trouvés entre les 2 génomes mais la structure en opérons n'est pas conservée.
- structure opéronique inconnue ou perdue: si aucun ou seulement un orthologue a été trouvé dans un opéron

(Itoh et al., 1999)

Génomes analysés

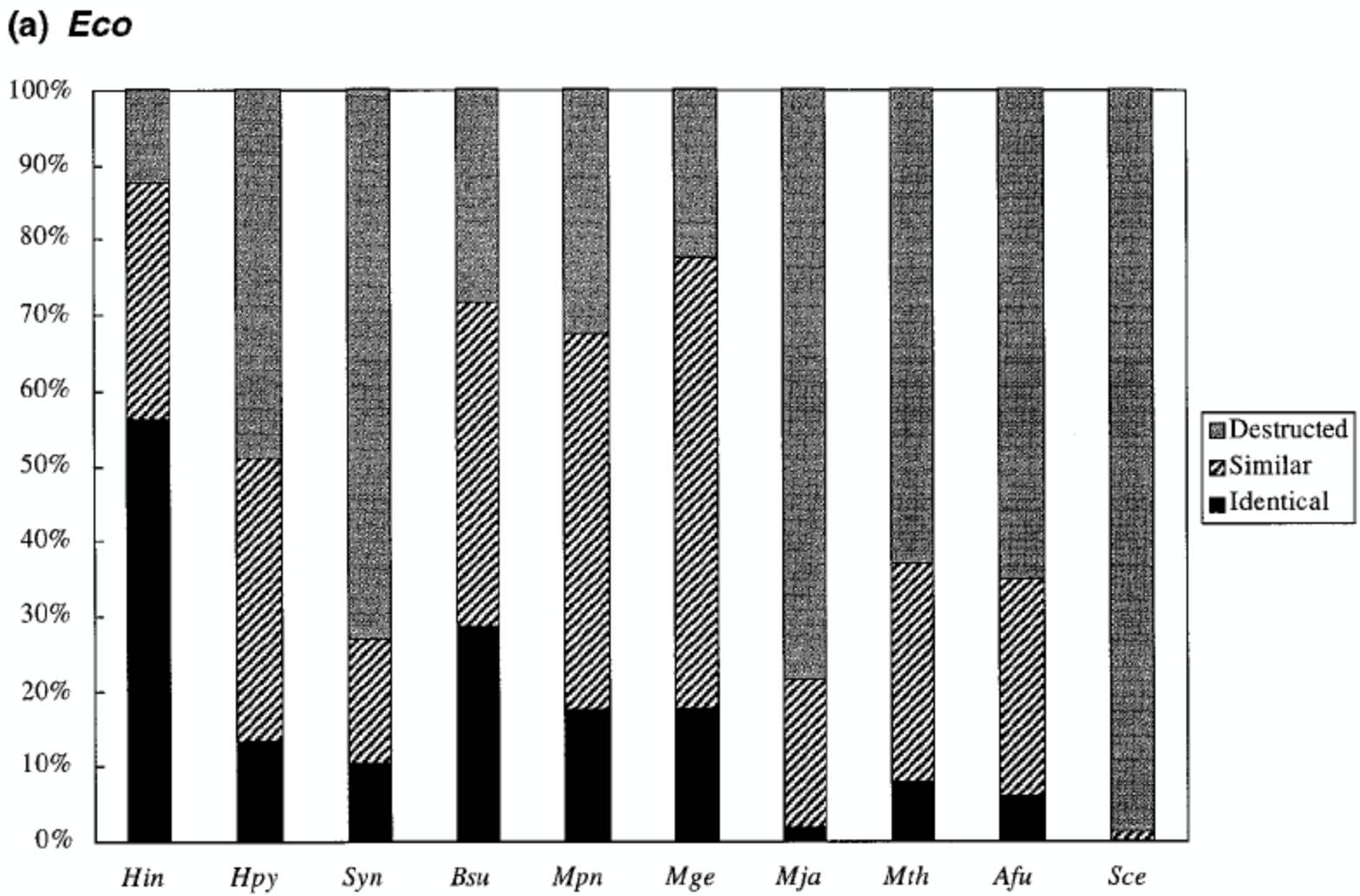
Arbre phylogénétique sans racine obtenu à partir de l'alignement des séquences protéiques EF-2/G (facteur d'élongation)

Eco: *E. coli*
Hin: *H. influenzae*
Hpy: *H. pyloru*
Bsu: *B. subtilis*
Mge: *M. genitalium*
Mpn: *M. pneumoniae*
Syn: *Synechocystis sp.*
Sce: *S. cerevisiae*
Mja: *M. jannaschii*
Mth: *M. thermoautotrophicum*
Afu: *A. fulgidus*
THICU: *Thiobacillus cuprinus*
ANANI: *Anacystis nidulans*
THETH: *Thermus aquaticus*
AQUAE: *Aquifex aeolicus*
METVA: *Methanococcus vannielii*
THEAC: *Thermoplasma acidophilum*
SCHPO: *S. pombe*
CHICK: *Gallis gallus*
DROME: *D. melanogaster*
CAEEL: *C. elegans*
HUMAN: *Homo sapiens*

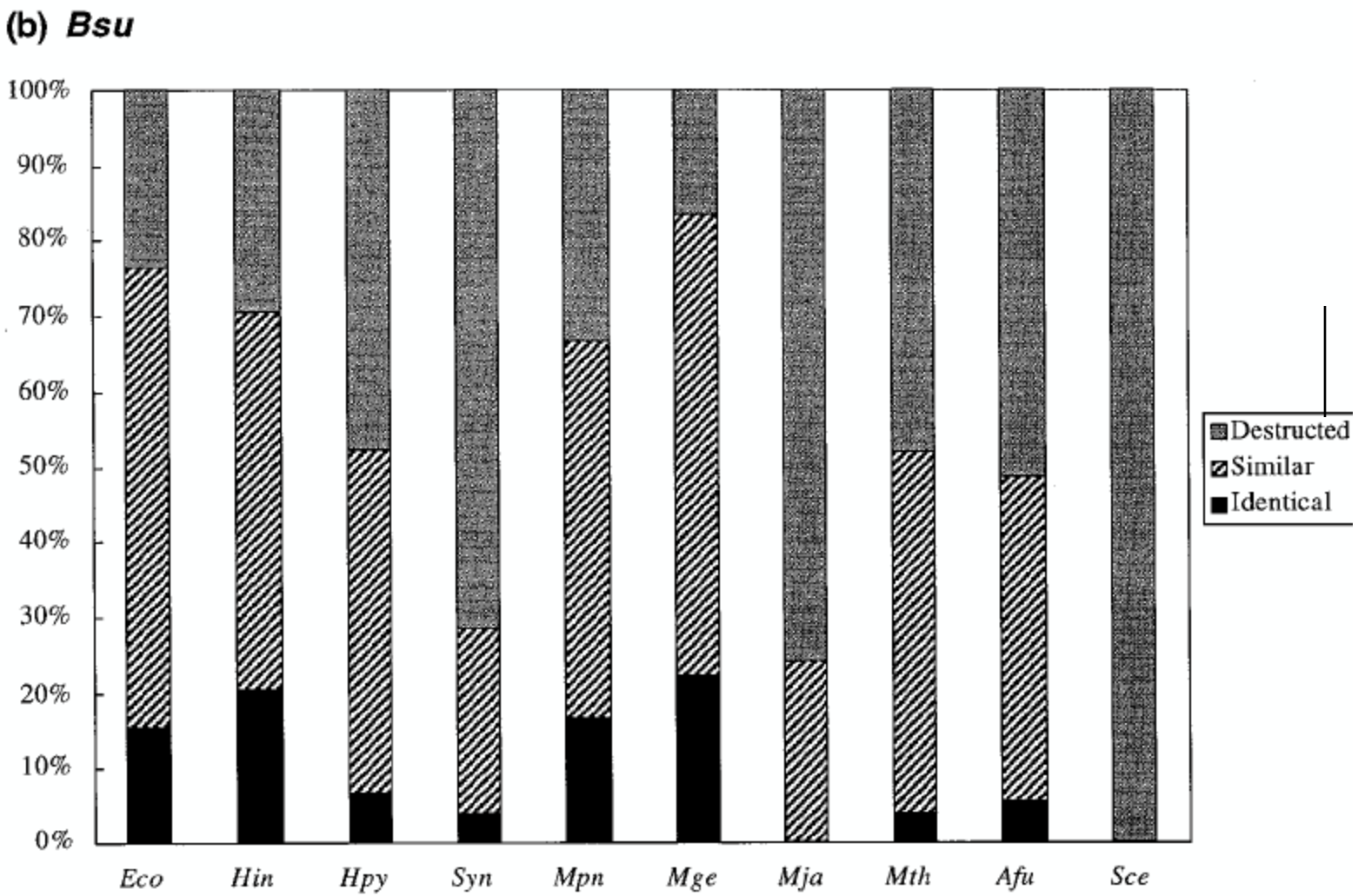


Les nombres correspondent aux valeurs de bootstrap pour 1000 répétitions.

Conservation de la structure en opérons au sein des génomes (Itoh et al., 1999)



Conservation de la structure en opérons au sein des génomes (Itoh et al., 1999)



Conservation de la structure en opérons au sein des génomes (Itoh et al., 1999)

Nombre D'IS dans les différents génomes analysés.

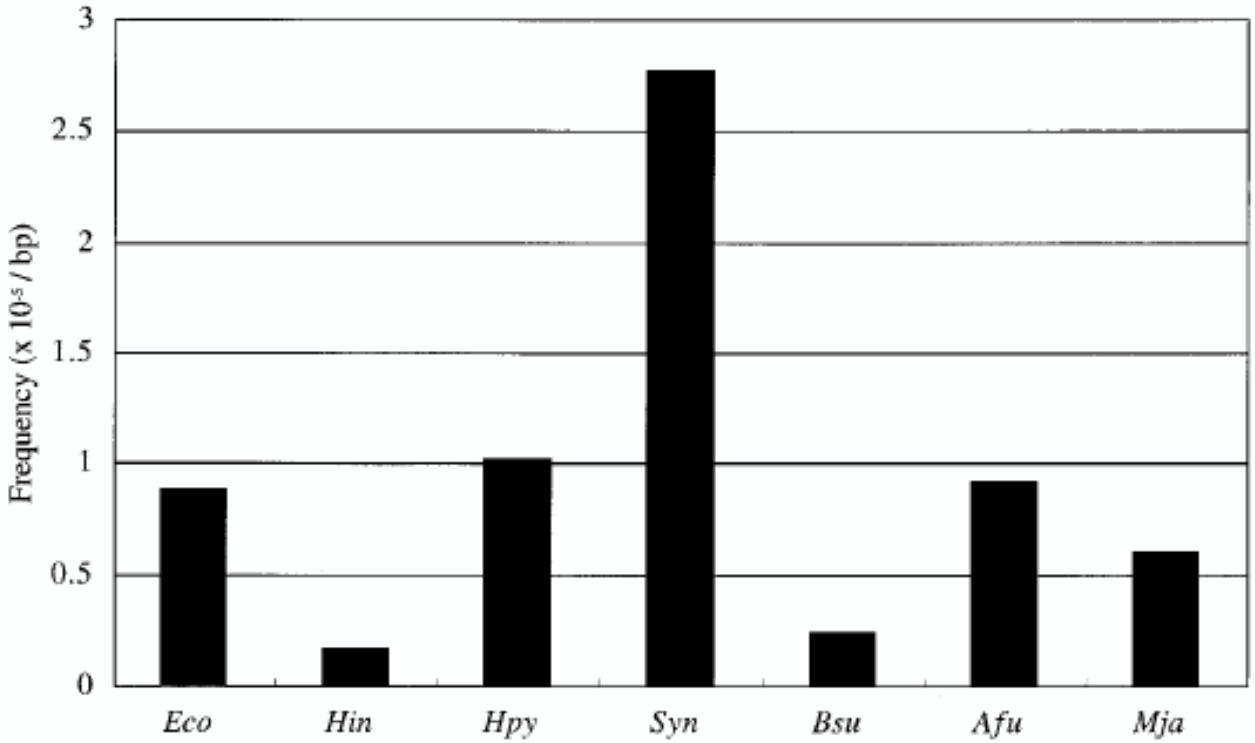


FIG. 8.—Frequency of occurrence of ISs per bp. The number of ISs in each genome was obtained from literature (Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997; Goffeau et al. 1997; Klenk et al. 1997; Kunst et al. 1997; Smith et al. 1997; Tomb et al. 1997). For *Syn,* the frequencies for transposases instead of ISs were calculated. Ten partial copies of ISs were included in *Hpy.* No IS-like element was reported in *Mge, Mpn,* or *Mth.*

Conservation de l'ordre des gènes chez les procaryotes

(Tamanes, 2001 http://genomebiology/2001/2/6/research)

Analyse de l'ordre des gènes pour répondre aux questions suivantes:

- La conservation est-elle la même chez tous les procaryotes ?
- Les régions conservées sont-elles distribuées uniformément sur le chromosome ?

Définitions opérationnelles:

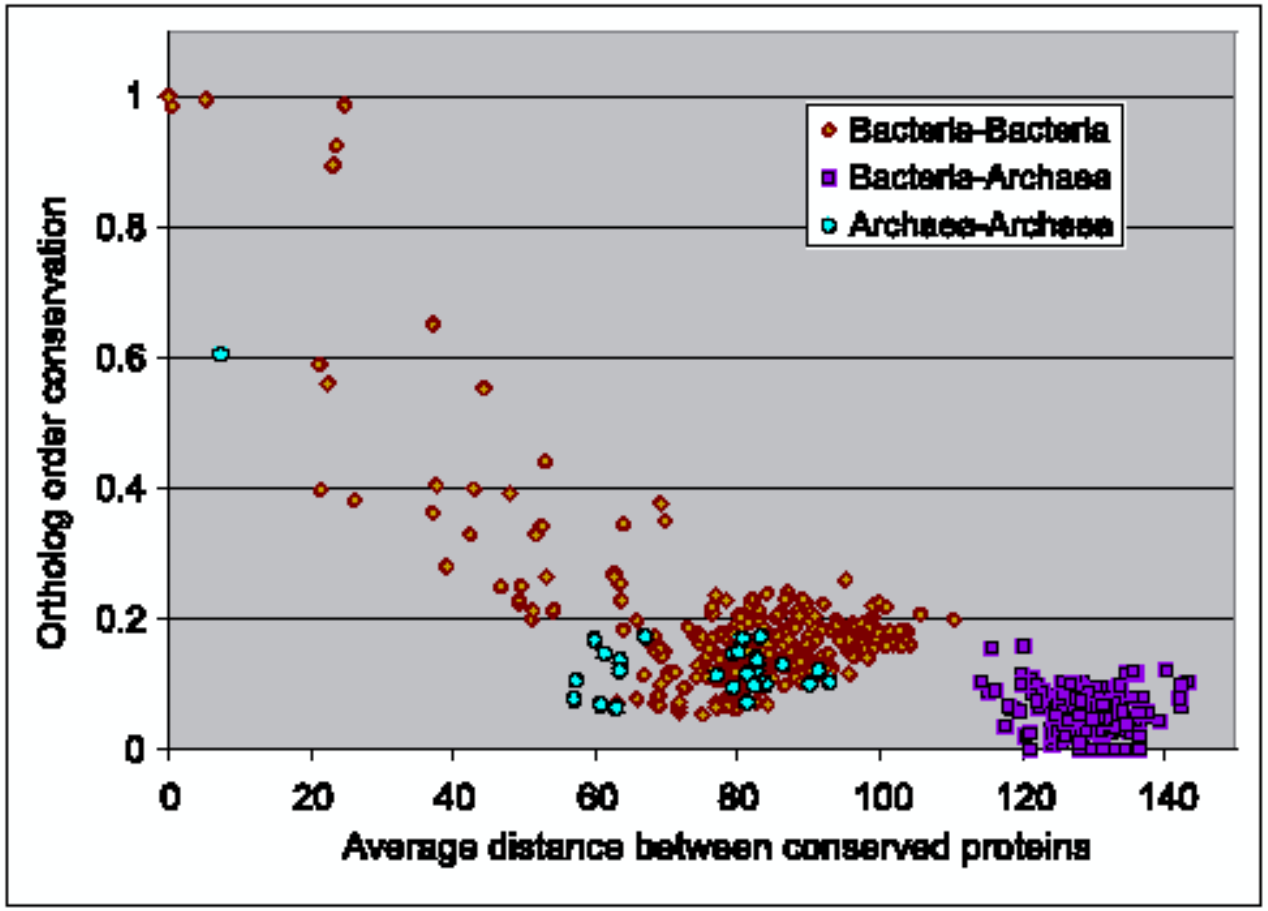
- 2 ORF sont dites homologues (recherche avec Blast) si :
  - l'alignement représente au moins 75% de la longueur des 2 ORF
  - la e-value est inférieur à 10^{-5}
- 2 ORF sont dites orthologues si on a une relation bijective de similarité ("best bidirectionnal hit" (BBH))
- La mesure de la conservation des gènes entre 2 génomes est donnée par le rapport :

$$\frac{\text{nombre de gènes localisés dans des suites conservées}}{\text{nombre total de gènes orthologues}}$$

- Une suite :
  - est constituée de gènes sur le même brin
  - contient au moins 3 gènes
  - au plus 3 insertions sont autorisées

Conservation de l'ordre des gènes dans les génomes procaryotes en fonction de la distance phylogénétique mesurée par la moyenne des distances obtenues par phylogénie moléculaire sur 24 protéines codées par des gènes conservés dans l'ensemble des organismes étudiés.

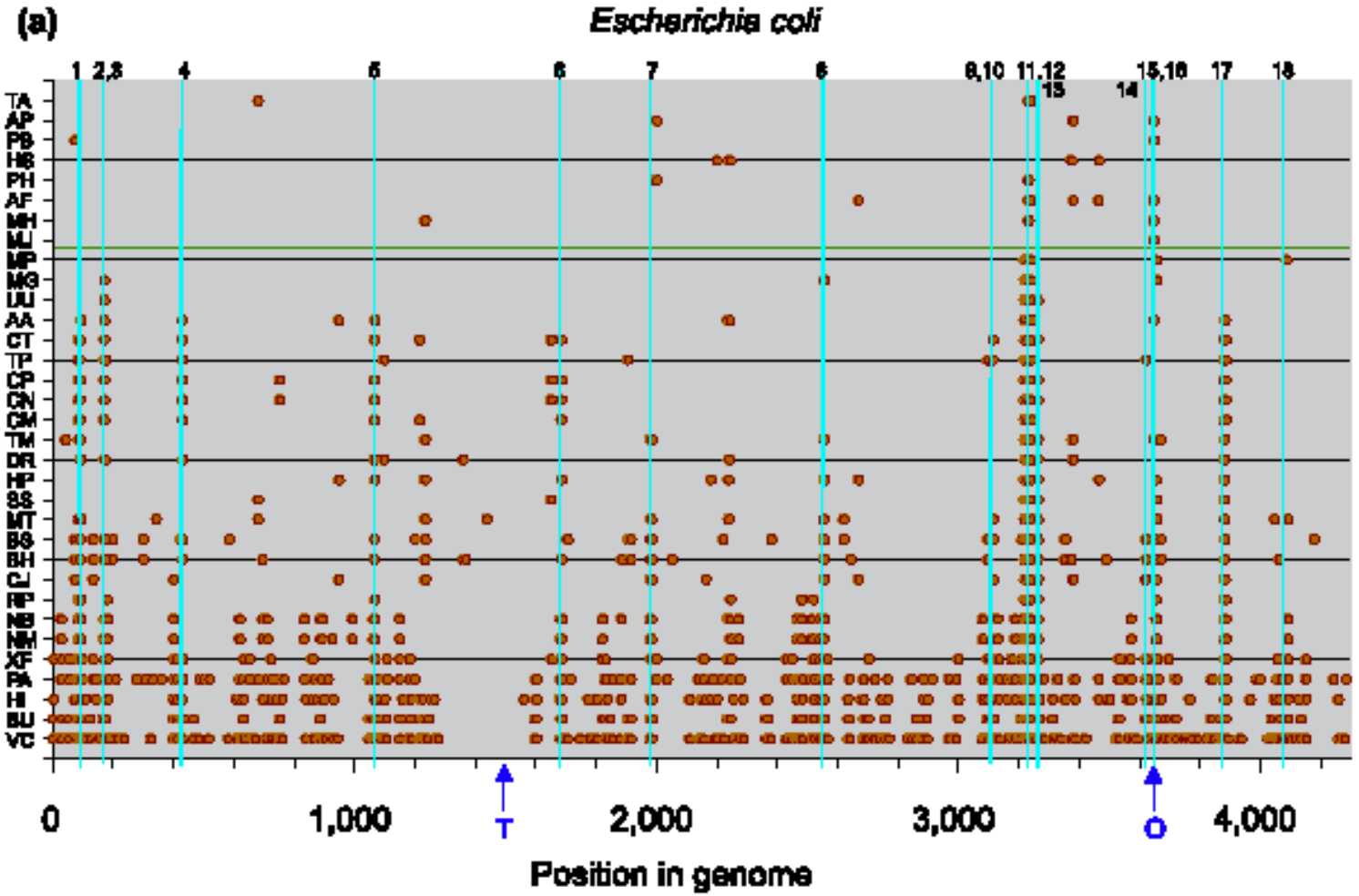
(Tamanes, 2001)



Distribution de la conservation des gènes le long des génomes (Tamanes, 2001)

Génome de référence: *E. coli*

Axe des *y*: génome individuel ordonné suivant la distance évolutive (SSU rRNA) croissante



T: Terminaison de réplication, O: origine de réplication

Analyses des suites de gènes conservées entre génomes

(Tamanes, 2001)

• Dans les suites de gènes conservées les produits de traduction sont généralement impliqués dans une même classe fonctionnelle.

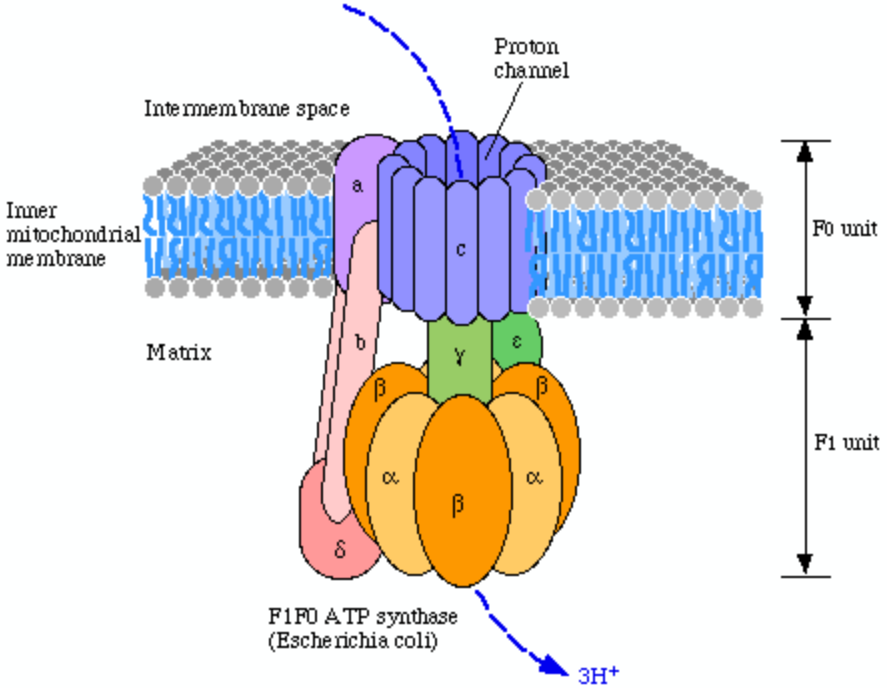
• Pas de classe fonctionnelle sur-représentée, à l'exception de la traduction à cause de l'opéron des protéines ribosomiques.

• Quand les suites sont impliquées dans le métabolisme, les gènes codent pour des enzymes qui agissent de façon séquentielle dans une voie métabolique ou qui forment des complexes enzymatiques.

Un exemple de suite de gènes conservés

Group 16

3652	*atpC*	24	Membrane-bound ATP synthase, F1 sector, epsilon-subunit	Energy metabolism
3653	*atpD*	48	Membrane-bound ATP synthase, F1 sector, beta-subunit	Energy metabolism
3654	*atpG*	52	Membrane-bound ATP synthase, F1 sector, gamma-subunit	Energy metabolism
3655	*atpA*	52	Membrane-bound ATP synthase, F1 sector, alpha-subunit	Energy metabolism
3656	*atpH*	39	Membrane-bound ATP synthase, F1 sector, delta-subunit	Energy metabolism
3657	*atpF*	30	Membrane-bound ATP synthase, F0 sector, subunit b	Energy metabolism
3659	*atpE*	30	Membrane-bound ATP synthase, F0 sector, subunit a	Energy metabolism



F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

Conservation de l'ordre des gènes : une empreinte des protéines qui interagissent physiquement

(Dandekar et al., 1998, TIBS, 23,324-328)

• Ordre des gènes est considérablement modifié quand l'identité des protéines orthologues de deux génomes est inférieure à 50%.

• Pour avoir une conservation de l'ordre des gènes significative du point de vue de l'évolution il faut donc étudier celle-ci dans des espèces éloignées.

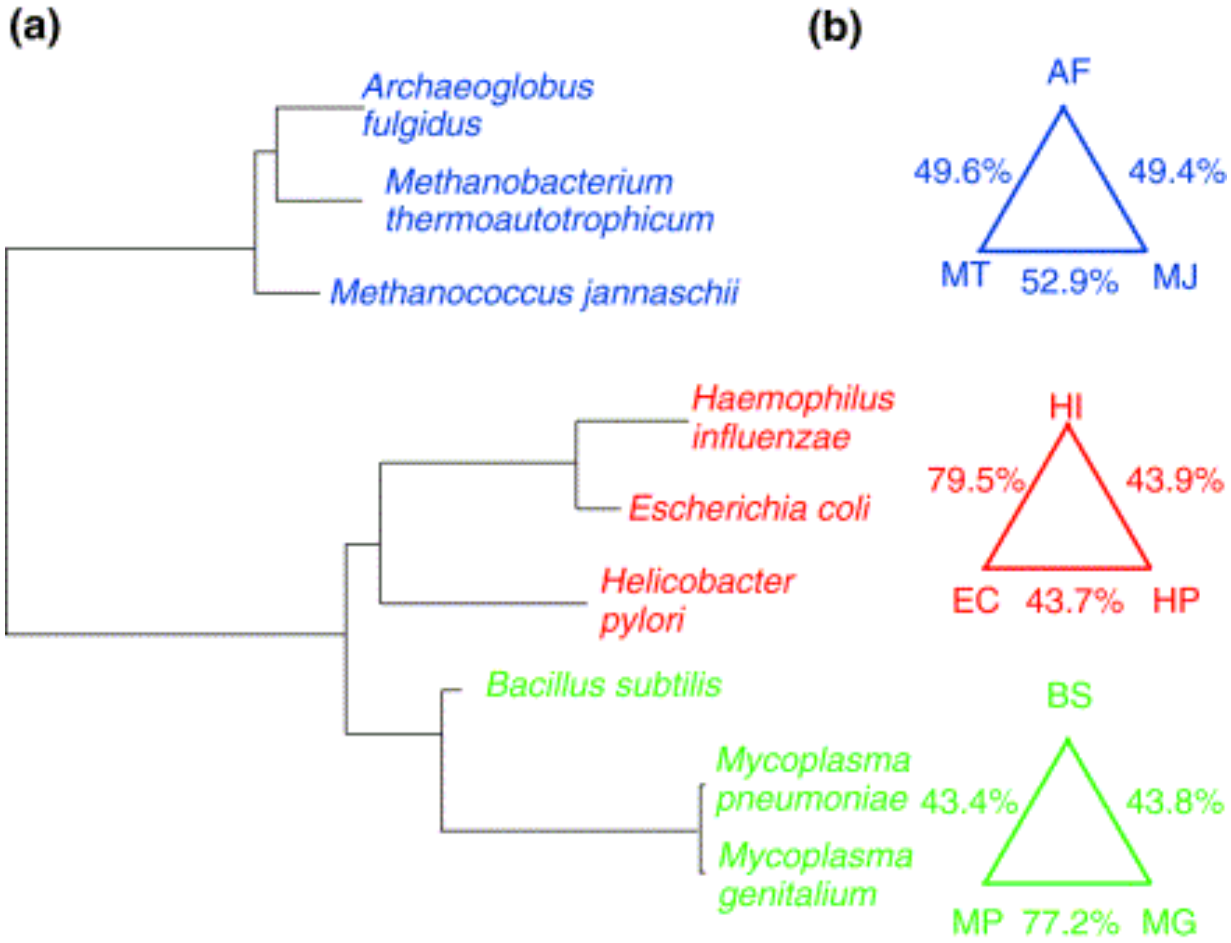
• Etude réalisée sur 3 groupes de 3 génomes avec 2 des distances intergénomiques ayant moins de 50% d'identité entre les orthologues communs.

- 3 protéobactéries (*E. coli, H. influenzae, H. pylori*)
- 3 bactéries Gram + (*B. subtilis, M. genitalium, M. pneumoniae*)
- 3 archébactéries (*M. jannaschii, M. thermoautotrophicum, A. fulgidus*)

• Pour éviter les problèmes de transferts horizontaux, seuls les gènes orthologues conservés dans le même ordre dans les trois génomes sont pris en compte.

Pourcentage d'identité entre les orthologues communs des deux espèces

(Dandekar et al., 1998)



Conservation de l'ordre des gènes : une empreinte des protéines qui interagissent physiquement

(Dandekar et al., 1998, TIBS, 23,324-328)

Résultats:

Environ 100 gènes orthologues sont conservés dans le même sens de transcription en paires ou en groupes dans les 3 génomes.

Parmi ces paires conservées:

- pour au moins 75%, les gènes codent pour des protéines dont les interactions physiques ont été démontrées expérimentalement:
  - protéines ribosomiques
  - sous unités de l'ARN polymérase
  - sous unités de l'ATP synthétase (cf. résultats de Tamanes)
  - certaines sous unités des transporteurs ABC
  - différentes sous unités enzymatiques
  - protéines de la division cellulaire (FtsA et FtsZ)

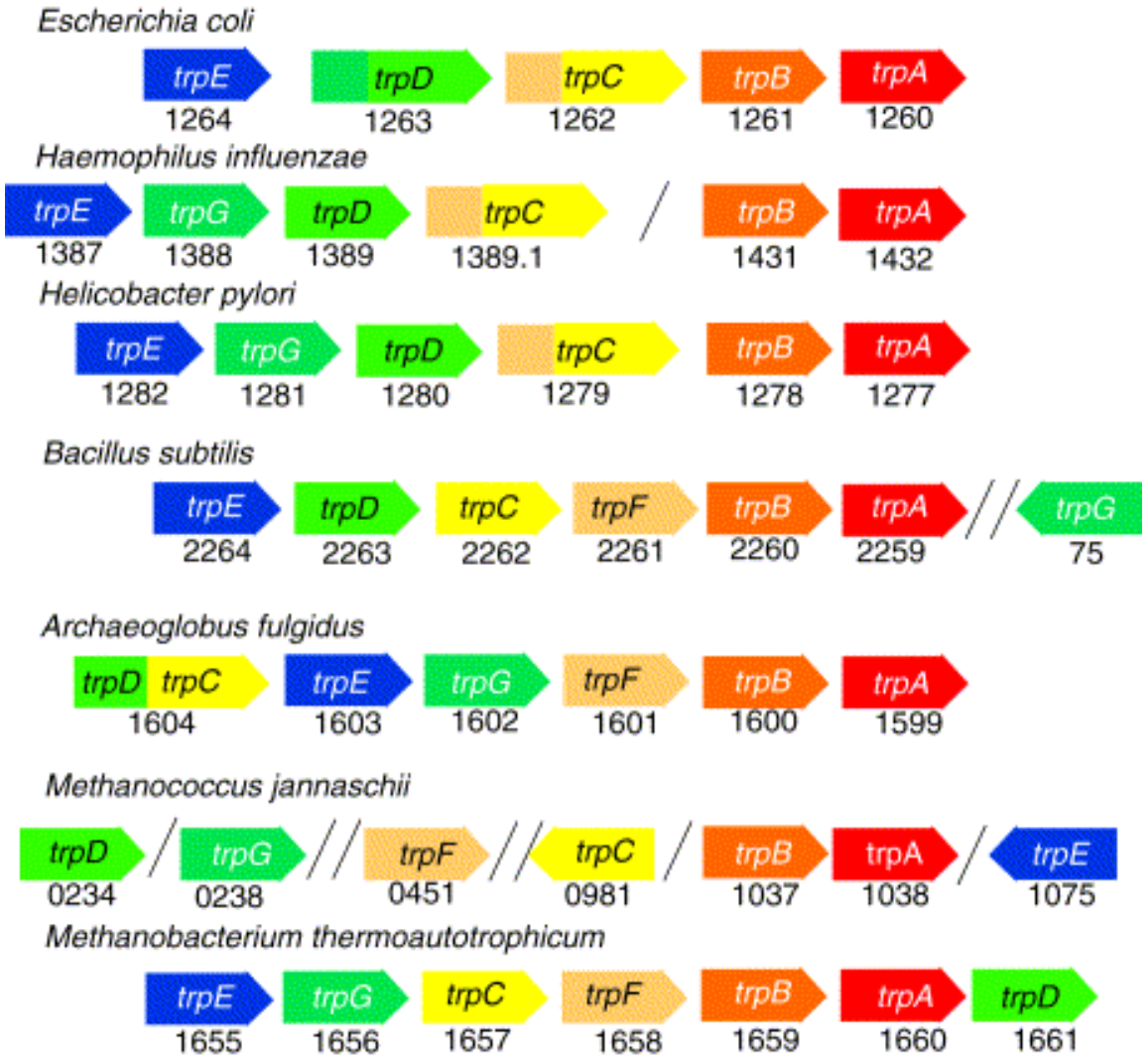
- pour 20%, de part la fonction des protéines codées on peut prédire qu'elles interagissent physiquement

- pour les 5% restants, soit pas de fonction décrite pour les protéines, soit pas d'évidence d'interaction entre-elles

Opéron tryptophane : organisation dans les 9 génomes comparés

(Dandekar et al., 1998)

Seule la paire de gènes *trpB-trpA* est conservée. Ces deux gènes codent pour deux sous unités de la tryptophane synthétase.



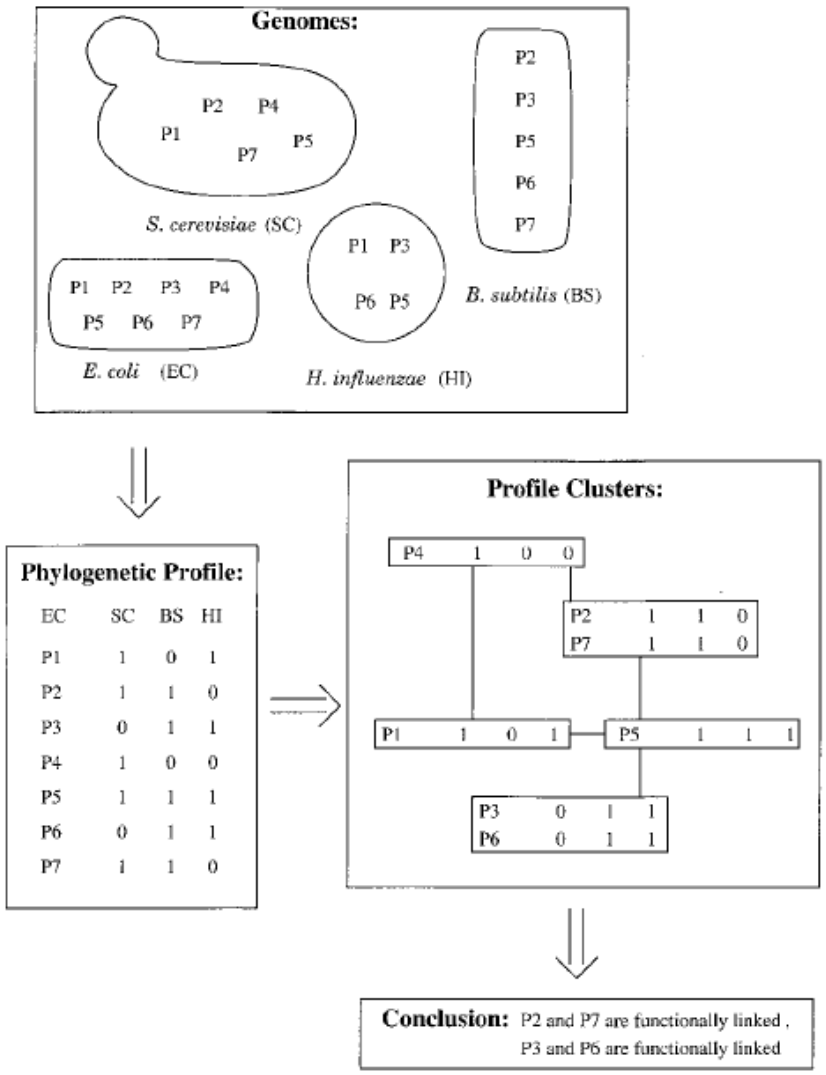
Profil phylogénétique

- Décrit la présence ou l'absence d'une protéine dans un ensemble de génomes de référence.
- Similarité entre profils phylogénétiques est un indicateur d'un couplage fonctionnel entre les produits des gènes (même pathway par exemple).
- Permet d'assigner des fonctions putatives à des protéines non caractérisées en se basant sur la similarité de leurs profils avec celui de protéines connues.

Basé sur l'hypothèse que les protéines qui font partie d'une même voie métabolique ou d'un même complexe physique vont probablement coévoluer et être cohéritées, si la fonction du système cellulaire auquel elles appartiennent est conservée au cours de l'évolution.

Profil phylogénétique

Méthode :



(extrait de Pellegrini *et al.*, 1999, PNAS,96,4285-88)

Pour chaque protéine d'intérêt :

- Recherche dans le groupe de génomes étudiés, les protéines orthologues (recherche des BBH)

En pratique :

- Créer une banque (*formatdb*) de toutes les séquences protéiques annotées dans les génomes.
- Utiliser la protéine d'intérêt comme query pour une recherche avec BlastP.
- Parser les résultats du Blast pour identifier les Best Bidirectional Hit

- Construire le profil phylogénétique qui est un vecteur contenant 1 si un orthologue est détecté dans le génome cible et 0 dans le cas contraire.

Mesure de la similarité entre profils :

- distance euclidienne
- coefficient de corrélation de Pearson
- distance de Hamming
- coefficient de Jaccard
- mutuelle information

Utilisation de méthodes de clustering pour regrouper les profils similaires

Profil phylogénétique

Distance de Hamming : permet de quantifier la différence entre deux séquences de symboles.

Cas des symboles binaires, alphabet *A* = {0,1}.

Soit *a* et *b* deux séquences de symboles binaires de longueur *n* avec,
a =(a_1,..,a_n) et b = (b_1,…,b_n),
la distance de Hamming entre *a* et *b* est : $d(a,b) = \sum_{i=1}^{n} (a_i \oplus b_i)$

$\oplus$: ou exclusif (XOR)

Exemple :

a = (1,1,0,0,1,0,1,1) et b = (0,0,0,1,1,1,1,1)

alors d(a,b) = 1+1+0+1+0+1+0+0 = 4

Table de vérité de XOR

A	B	R = A⊕B
0	0	0
1	0	1
0	1	1
1	1	0

Profil phylogénétique

Coefficient ou indice de Jaccard: permet de mesurer la similarité entre des échantillons. C'est le rapport entre la cardinalité (la taille) de l'intersection des ensembles sur la cardinalité de l'union des ensembles.

$$J(A,B) = \frac{|A \cup B|}{|A \cap B|}$$

Similarité entre des ensembles binaires A et B avec $A = (a_1,..,a_n)$ et $B = (b_1,\ldots,b_n)$.
On peut définir plusieurs quantités caractérisant les deux ensembles :

M_{11} = nombre d'attributs qui valent 1 dans A et B
M_{10} = nombre d'attributs qui valent 1 dans A et 0 dans B
M_{01} = nombre d'attributs qui valent 0 dans A et 1 dans B
M_{00} = nombre d'attributs qui valent 0 dans A et dans B
Avec $M_{11} + M_{10} + M_{01} + M_{00} = n$

L'indice de Jaccard devient : $$J = \frac{M_{11}}{M_{11} + M_{01} + M_{10}}$$

Exemple : A = (1,1,0,0,1,0,1,1) et B = (0,0,0,1,1,1,1,1)

$M_{11} = 3$ $M_{10} = 2$ $M_{01} = 2$ $M_{00} = 1$ on a J = 3/7 = 0.4286

Profil phylogénétique

Méthode plus sophistiquée :

- à la place d'un vecteur de 1 et 0, vecteur contenant un score calculé comme suit permettant de capturer différents degrés de divergence de séquences :

pour chaque protéine P_i et son hit de plus fort score dans le génome j, la valeur dans le vecteur p_{ij} est donnée par :

$$p_{ij} = \frac{1}{\log E_{ij}}$$

Avec E_{ij} correspondant à la e-value de Blast
Et $p_{ij} = 1$ si $p_{ij} > 1$

Protein	*EC*	*SC*	*CE*	*DM*	*AT*
>P1	1.0	1.0	1.0	0.0	0.0
>P2	0.0	1.0	0.8	0.6	1.0
>P3	0.0	1.0	0.8	0.6	1.0

1.0 (absence) → 0.0 (presence)

Profil phylogénétique

- Calcul de l'information mutuelle

L'utilisation de cette mesure est recommandée pour pouvoir capturer des relations linéaires mais aussi non linéaires présentes dans l'ensemble de données. Elle mesure la dépendance mutuelle entre deux variables.

L'information mutuelle *MI* entre deux protéines A et B est calculée comme :

$$MI(AB) = H(A) + H(B) - H(A,B)$$

où

$$H(A) = -\sum p(a) \ln p(a)$$

H(A) représente l'entropie marginale de la distribution de la probabilité p(a) du gène A dans chaque génome de référence. (Mesure l'incertitude associée à une variable aléatoire)

$$H(A,B) = -\sum\sum p(a,b) \ln p(a,b)$$

H(A,B) représente l'entropie (incertitude) restante d'une variable aléatoire A étant donné la valeur d'une seconde variable aléatoire B

En pratique, l'information mutuelle est calculée à partir d'un histogramme des valeurs des p_{ij} déterminées précédemment construit avec des intervalles de 0.1. Le calcul des valeurs utilise plutôt le log_2 que le *ln*.

Profil phylogénétique

(Extrait du chapitre 9 de Bioinformatics, volumeII: Structure, Function and Applications, vol. 453, Humana Press, 2008)

Calcul de la mutuelle information en pratique

Bins	P1	P2	P3
0	2	1	1
0.1	0	0	0
0.2	0	0	0
0.3	0	0	0
0.4	0	0	0
0.5	0	0	0
0.6	0	1	1
0.7	0	0	0
0.8	0	1	1
0.9	0	0	0
1	3	2	2

$p(P1_{Bin0.0}) = 2/5 = 0.4$ et $-(p(P1_{Bin0.0}) \ln p(P1_{Bin0.0})) = -0.3665$

H(P1) est la somme des valeurs de toutes les classes (Bins) pour la protéine P1 .

Pour calculer H(P1,P2)

- on va calculer le nombre de fois où les protéines apparaissent conjointement dans les deux classes (Bins) considérées

`Bin(1.0,0.0) =1; Bin(1.0,0.1)=0 etc.`

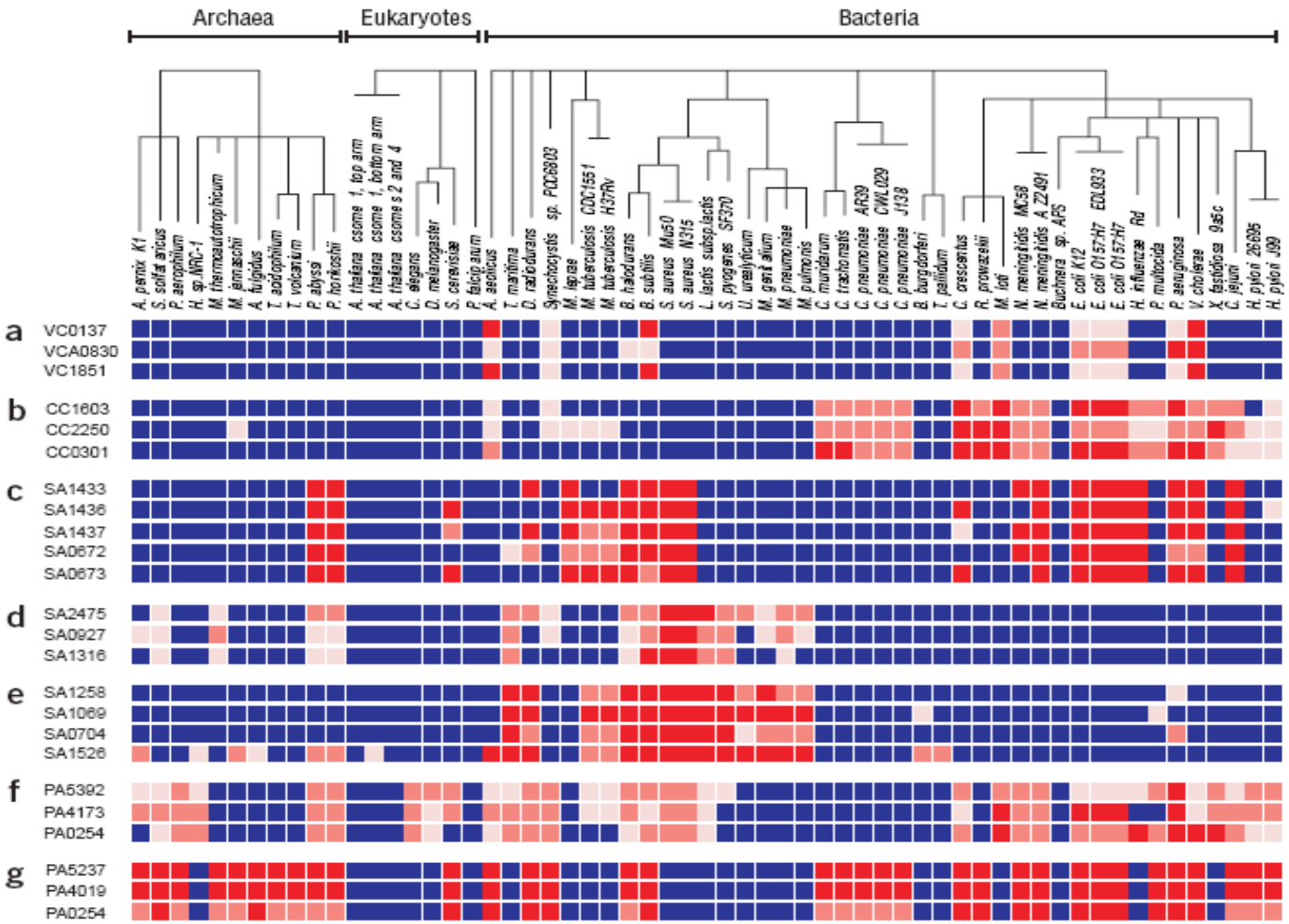
Nombre de fois où on a conjointement P1 dans la classe 1.0 et P2 dans la classe 0.0 etc.

- on va en déduire pour chaque classe (Bin) la probabilité H(P1,P2)

$p(P1,P2)_{Bin(1.0,0.0)} = 1/5 = 0.2$ etc

- La valeur finale de H(P1,P2) est obtenue en sommant les valeurs obtenues pour chaque classe (Bin)

Profil phylogénétique



<u>Représentation de profils phylogénétiques</u>

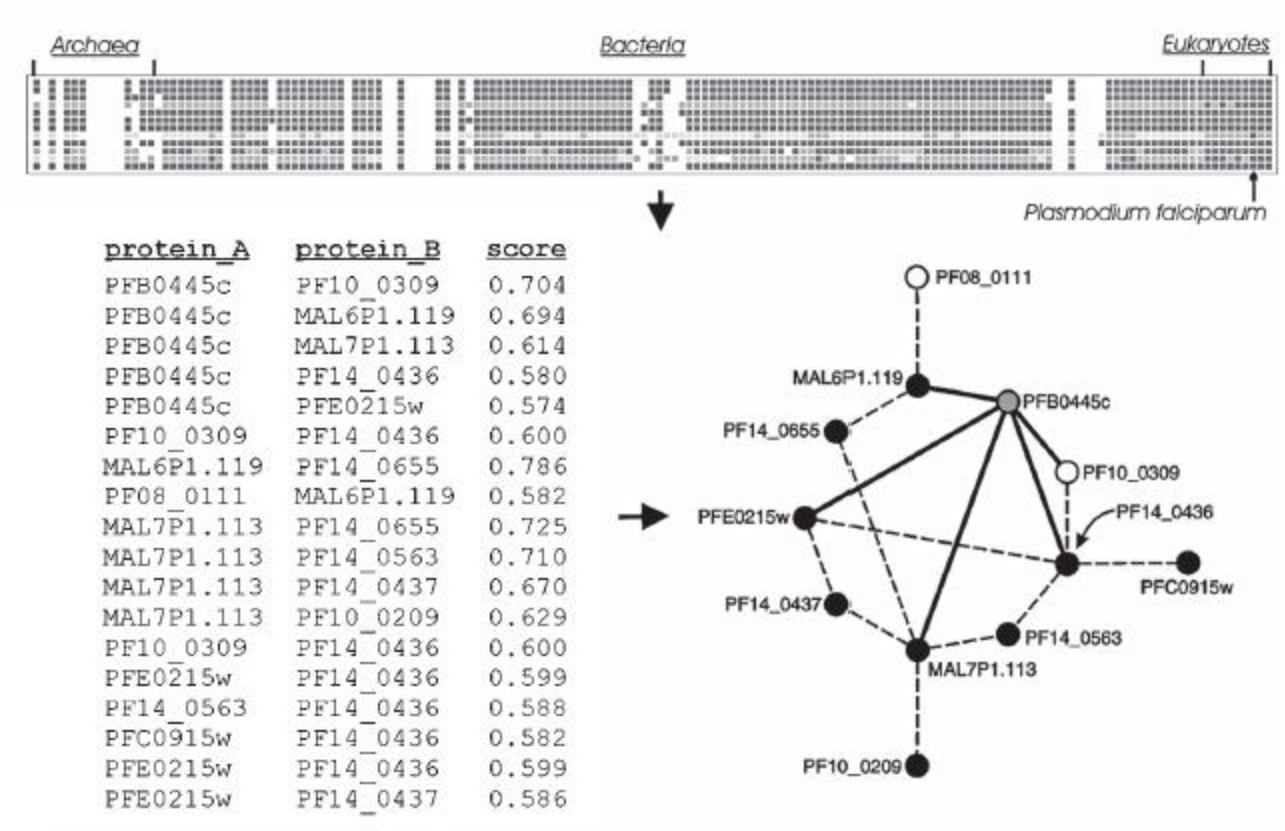
Intensité du rouge dénote l'intensité de la similarité de séquence entre la query (gauche) et le best hit trouvé dans le génome.

Bleu absence de similarité.

(Exemple tiré de Date et Marcotte (2003) Nat. Biotechnology ,21 : 1055-62)

Profil phylogénétique

(Extrait du chapitre 9 de Bioinformatics, volumeII: Structure, Function and Applications, vol. 453, Humana Press, 2008)



protein_A	protein_B	score
PFB0445c	PF10_0309	0.704
PFB0445c	MAL6P1.119	0.694
PFB0445c	MAL7P1.113	0.614
PFB0445c	PF14_0436	0.580
PFB0445c	PFE0215w	0.574
PF10_0309	PF14_0436	0.600
MAL6P1.119	PF14_0655	0.786
PF08_0111	MAL6P1.119	0.582
MAL7P1.113	PF14_0655	0.725
MAL7P1.113	PF14_0563	0.710
MAL7P1.113	PF14_0437	0.670
MAL7P1.113	PF10_0209	0.629
PF10_0309	PF14_0436	0.600
PFE0215w	PF14_0436	0.599
PF14_0563	PF14_0436	0.588
PFC0915w	PF14_0436	0.582
PFE0215w	PF14_0436	0.599
PFE0215w	PF14_0437	0.586

Comparaison deux à deux des profils (calcul de la mutuelle information (score)) . Les liens primaires (protéines liées directement à la protéine query (gris) et les liens secondaires (protéines liées aux protéines cœurs (core) révèlent des sous-réseaux et permettent d'assigner une fonction potentielle aux protéines non caractérisées fonctionnellement (blanche).