Méthode du maximum de vraisemblance

Méthode introduite en reconstruction phylogénétique par Neyman (1971) et surtout Felsenstein (1981), premier à avoir développé un algorithme efficace applicable aux séquences nucléotidiques. Cependant, cette méthode, nécessitant l'estimation de nombreux paramètres, est très gourmande en temps calcul et ce n'est qu'en 1990 qu'elle a commencé à être appliquée aux séquences protéiques grâce aux capacités accrues des ordinateurs. On ne pouvait cependant l'appliquer que sur un petit nombre de séquences. Ce n'est que récemment (2003) que cette méthode peut être appliquée à un grand jeu de données grâce au développement de programmes rapides et aux performances toujours croissantes des ordinateurs.

Maximum de vraisemblance

Les objets étudiés en phylogénie moléculaire, les séquences, sont le résultat d'une histoire évolutive qui nous est inconnue mais que l'on peut essayer de reconstruire sachant que cette histoire intègre plusieurs composantes :

- les relations de parenté entre les séquences représentées par la topologie *t* de l'arbre.
- la quantité d'évolution qui s'est écoulée entre chacune des lignées étudiées et qui est représentée par l'ensemble des longueurs des branches *bi* .
- le processus qui gouverne l'évolution de ces séquences, le modèle évolutif considéré composé lui-même d'un certain nombre de paramètres θ.

Les valeurs des différents paramètres ne sont que très rarement connues.
On va donc devoir estimer, en fonction des données actuelles, *i.e.*, les séquences, cet ensemble Θ de paramètres (la topologie *t*, les longueurs de branches b_i et les paramètres θ du modèle évolutif).

On a un grand nombre de scénarios évolutifs possibles. Cependant certains d'entre eux sont plus susceptibles que d'autres de produire les séquences actuelles.

Le but des méthodes de maximum de vraisemblance est d'identifier ces scénarios, c'est-à-dire de trouver les valeurs des paramètres de Θ qui maximisent la probabilité d'observer les séquences actuelles.

Maximum de vraisemblance

Hypothèses

- Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
- Les sites évoluent indépendamment les uns des autres (restrictive).
- Les sites évoluent selon la même loi (on peut affaiblir cette hypothèse).
- Les taux de substitution ne changent pas au cours du temps le long d'une branche. Ils peuvent varier entre branches, c'est-à-dire, que l'évolution des séquences est indépendante d'une lignée à l'autre.

Deux applications du maximum de vraisemblance en phylogénie :

- Estimer la vraisemblance d'un ensemble d'hypothèses.
- Rechercher parmi tous les ensembles Θ de valeurs de paramètres possibles celui qui possède la vraisemblance la plus élevée. Comme on a vu que la topologie *t* faisait partie de ces paramètres, cela permet de rechercher l'arbre qui possède la plus forte vraisemblance étant donné la valeur des autres paramètres.

Maximum de vraisemblance

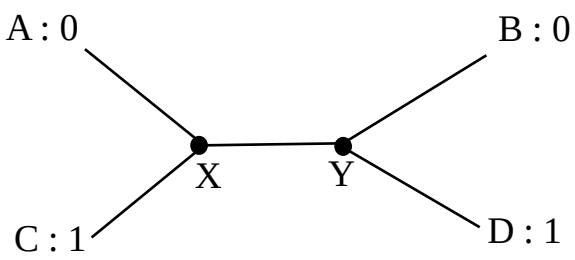
Etant donné un modèle d'évolution trouver l'arbre maximisant la probabilité d'obtenir les séquences actuelles.

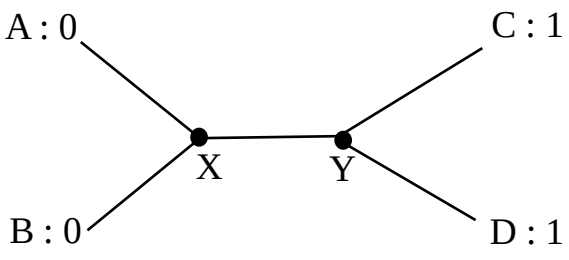
Considérons un exemple simple :

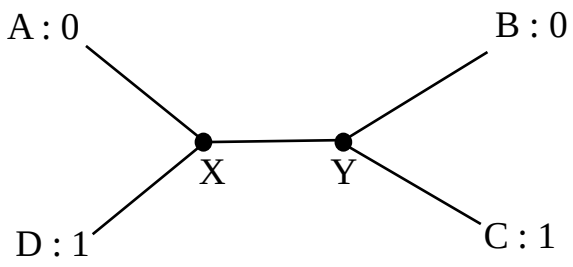
- 4 OTU (A, B, C, D) et un caractère qui peut avoir deux états 0 ou 1.
- On observe l'état de caractère 0 pour A et B et l'état 1 pour C et D.
- On a le modèle d'évolution suivant :

	0	1
0	0.8 (4/5)	0.2 (1/5)
1	0.2 (1/5)	0.8 (4/5)

Il existe trois arbres possibles avec une arête interne :

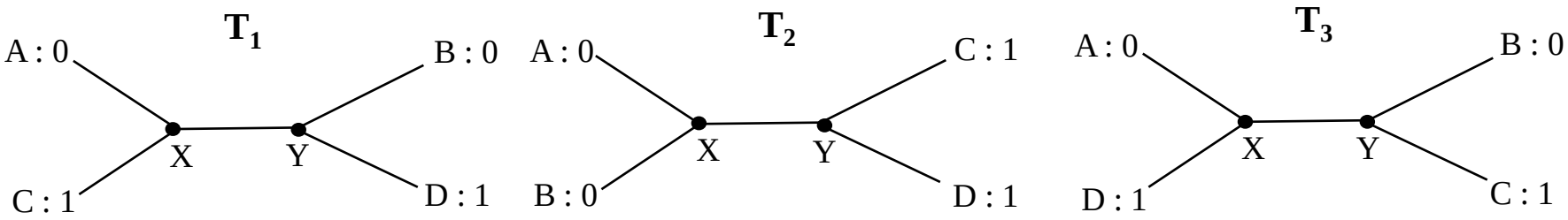






Maximum de vraisemblance

On va calculer la vraisemblance L(T) (Likelihood) de chaque topologie :



Les topologies T_1 et T_3 sont équivalentes, donc on calcule uniquement L(T1).

X	Y	$L(T_1)$	$L(T_2)$
0	0	$(4/5)^3(1/5)^2$	$(4/5)^3(1/5)^2$
0	1	$(4/5)^2(1/5)^3$	$(4/5)^4(1/5)^1$
1	0	$(4/5)^2(1/5)^3$	$(4/5)^0(1/5)^5$
1	1	$(4/5)^3(1/5)^2$	$(4/5)^3(1/5)^2$
Somme		32/625=0.0512	77/625=0.1232

L'arbre T_2 est plus vraisemblable que l'arbre T_1

Maximum de vraisemblance

Le cas d'une topologie calculée sur des séquences nucléiques :

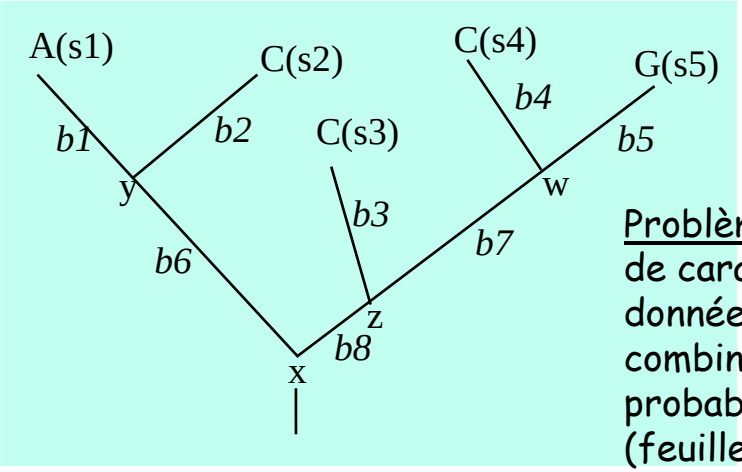
Première étape :

- calculer la vraisemblance à un site quelconque, c'est-à-dire la probabilité que les hypothèses soient à l'origine des états de caractères observés à ce site.

Soit un ensemble de données D composé par :

un arbre enraciné donné obtenu sur 5 séquences nucléiques
un site donné *i*
un jeu de longueurs des branches

Hypothèse T: les bases observées à ce site *i* ont évolué le long de cet arbre.
Vraisemblance de T par rapport à D(i) : probabilité que les données correspondent à l'hypothèse



$$L(T) = Prob(D(i)|T)$$

Problème : pour calculer cette probabilité, il faut connaître les états de caractères présents aux nœuds internes et à la racine. Or ces données sont inconnues. Il faut donc évaluer, pour chaque combinaisons d'états de caractères présents aux nœuds internes, la probabilité qu'ils aient conduits aux bases actuelles observées (feuilles). Par exemple, quelle est la probabilité d'observer A(s1), C(s2), C(s3), C(s4) et G(s5) sachant que x = A, y = A, z = A et w = A.

Maximum de vraisemblance

Si nous prenons maintenant en compte les *m* sites alignés, comme nous avons comme hypothèse que les sites évoluent de façon indépendante, la vraisemblance de la topologie par rapport aux données D est donnée par le produit des vraisemblances calculées pour chaque site :

$$L = \Pr ob(D \mid T) = \prod_{i=1}^{m} \Pr ob(D_{(i)} \mid T)$$

La valeur de L correspond à la probabilité que les séquences aient évoluées d'après l'arbre T.

Maximum de vraisemblance

- C'est la méthode la mieux justifiée au plan théorique.
- Des expériences de simulation de séquences ont montré que cette méthode est supérieure aux autres dans la plupart des cas.
- Mais c'est une méthode très lourde en calculs.
- Il est presque toujours impossible d'évaluer tous les arbres possibles car ils sont trop nombreux. Une exploration partielle de l'ensemble des arbres est réalisée en utilisant des méthodes de réarrangements locales ou globales similaires à celles utilisées en parcimonie.

- Deux méthodes principales : TREE-PUZZLE et PhyML.

Robustesse des topologies

Les arbres phylogénétiques construits avec une méthode quelconque ne sont qu'une estimation de l'histoire évolutive des séquences. Il est donc important de disposer de méthodes permettant d'évaluer statistiquement cet estimateur qu'est l'arbre. Les méthodes les plus couramment utilisées seront décrites.

Robustesse des topologies

Bootstrap et Jackknife :

Deux méthodes basées sur des techniques de ré-échantillonage.
Méthodes empiriques permettant d'inférer la variabilité des paramètres quand les modèles sont trop complexes pour pouvoir en calculer la variance.
Introduites en phylogénie par Felsenstein (1985) .
Elles sont basées toutes les deux sur l'hypothèse que l'évolution des sites est indépendante.

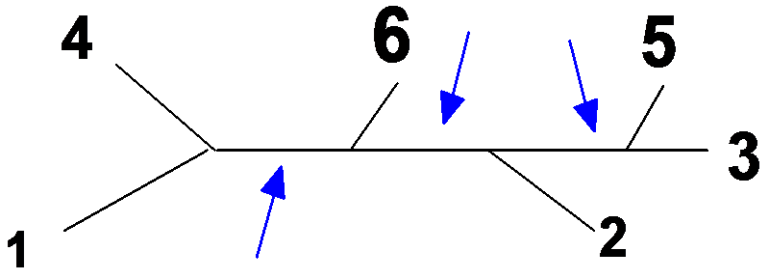
➡ Si un arbre est robuste, c'est-à-dire fortement soutenu par les données, alors sa variance sera faible.

➡ Si un arbre est peu robuste alors il présentera une grande variabilité

Le bootstrap non paramétrique (appelé couramment bootstrap) est la méthode la plus couramment utilisée pour mesurer les incertitudes sur les arbres.
Peut être utilisée en combinaison de n'importe quelle méthode de reconstruction.

Le bootstrap

- L'information phylogénétique contenue dans un arbre non raciné réside entièrement dans ses branches internes.

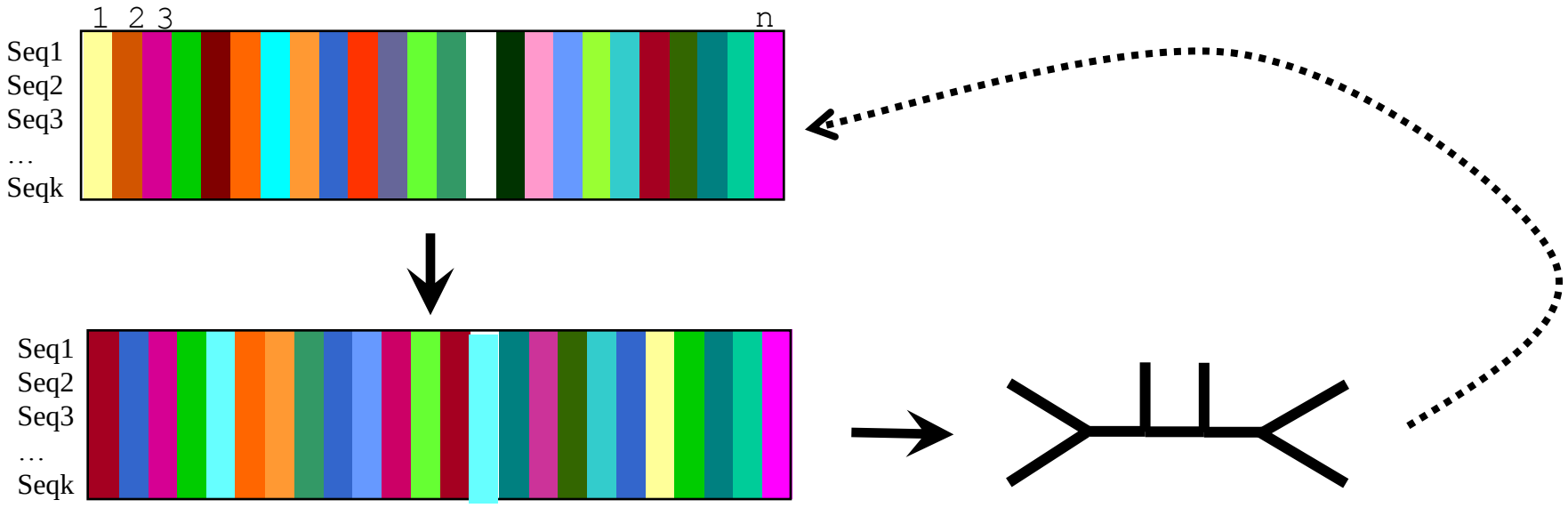


internal branch separating group 1+4
from group 6+2+5+3

- La forme de l'arbre est déterminée par la liste des branches internes.
- Evaluer la fiabilité d'un arbre = évaluer celle de chaque branche interne.

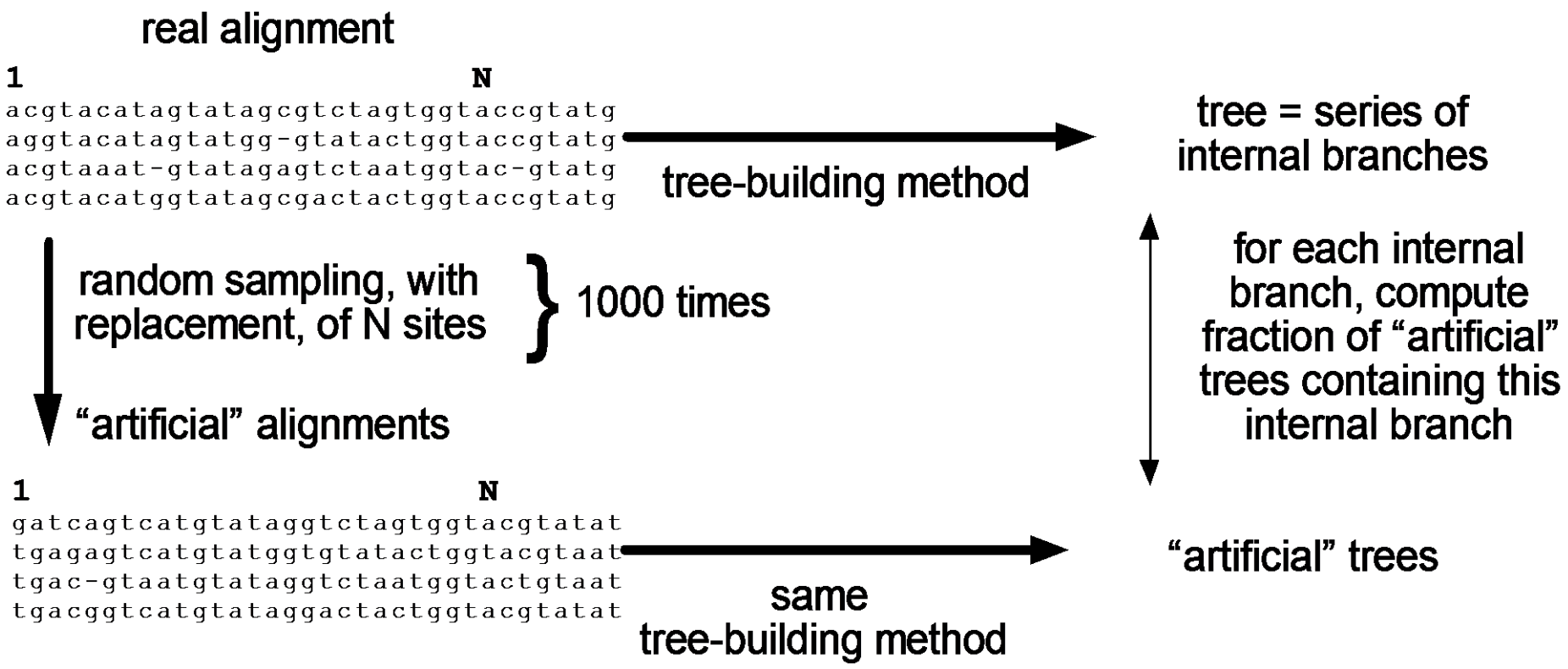
Le bootstrap : procédure

- Tirage avec remise de *n* positions parmi *n* positions
  - Le tirage avec remise de positions, en respectant l'effectif original, revient à conférer un poids aléatoire aux positions
- Construire l'arbre phylogénétique
- Répéter 1) et 2) un grand nombre de fois (1000)



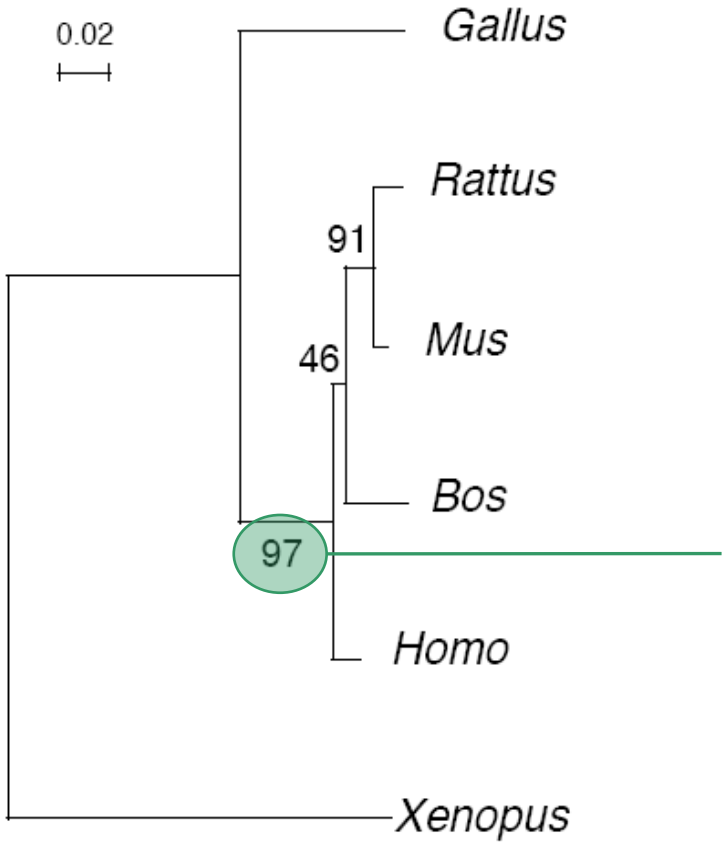
Pour chaque arbre reconstruit on compte le nombre de fois que l'on observe la branche interne. Le soutien de la branche est exprimé en pourcentage de réplications. Si la branche est observée dans tous les arbres, la valeur du bootstrap est égale à 100.

Le bootstrap : procédure



Le bootstrap

Test individuellement la validité de chaque branche interne de l'arbre. Pour cela, on calcule le pourcentage de fois où chaque branche interne de l'arbre de départ se retrouve dans les arbres construits par rééchantillonage. Ce pourcentage correspond à la valeur du bootstrap.



Estimation statistique de la **confiance** à accorder à une branche

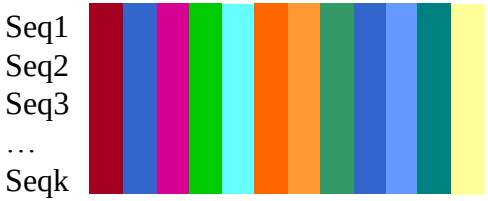
97% des 1000 arbres contenaient cette branche

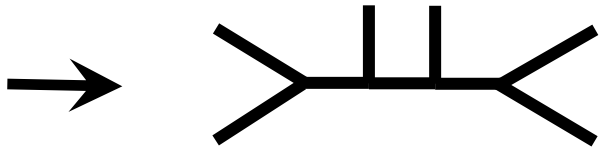
Le jackknife

Principe : Tirage aléatoire sans remise afin de construire les échantillons. En pratique, approche la plus utilisée, construire des échantillons de taille n/2, c'est-à-dire contenant la moitié des sites de l'alignement.

Alignement de départ

Alignement après tirage de n/2 sites





- Comme pour le bootstrap, une fois les X réplicats construits, les arbres correspondant à chaque réplicat sont calculés.

- La mesure de la robustesse de l'arbre se fait également en calculant le pourcentage de fois où chaque branche interne de l'arbre de départ est retrouvée dans les arbres issus du ré-échantillonage.

- Donc bootstrap et jackknife utilisent des techniques très similaires.

Le bootstrap : interprétation

Problème beaucoup discuté .

De manière générale, une faible valeur de bootstrap indique que la quantité d'information supportant la bipartition induite par une branche interne est faible.

Quel seuil ?

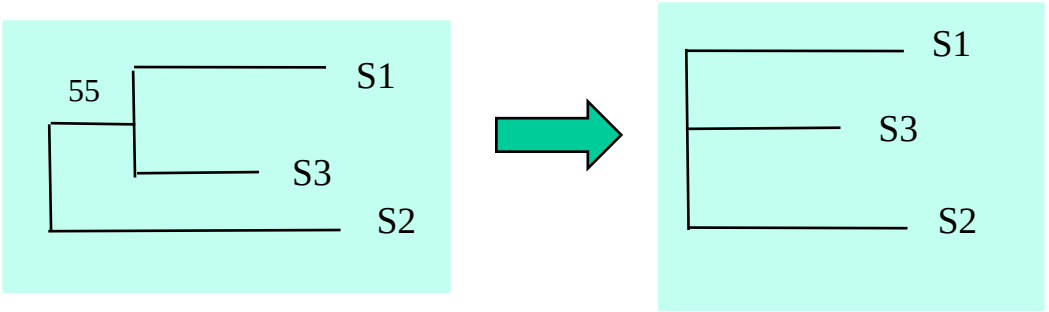
Si on applique les critères standards utilisés en statistique, il ne faudrait considérer comme valide que les branches ayant un support de bootstrap ≥ 95% (sinon la branche n'existe pas).
Des travaux ont montré que ce seuil était trop élevé, notamment ceux de Hillis et Bull (1993, Syst. Biol., 42, 182-92) qui à l'aide de simulations ont montré que des supports de 70% pouvaient correspondre à des groupements significatifs.
Cependant résultat pas généralisable à toutes les analyses.

La procédure de bootstrap n'aide pas à déterminer si la méthode de construction d'arbre est bonne. Un arbre faux peut avoir un score de bootstrap de 100 % pour chacune de ses branches !

Le bootstrap : interprétation

Remplacer par des multifurcations.

Quand faibles valeurs de bootstrap, possible de remplacer les branches incriminées par des multifurcations indiquant que les données ne permettent pas de résoudre sans ambiguïté l'ordre d'émergence des différentes lignées.



L'attraction des longues branches

- Toutes les méthodes sont sujettes à l'artéfact dit d'attraction des branches longues, mais la parcimonie y est plus sensible.

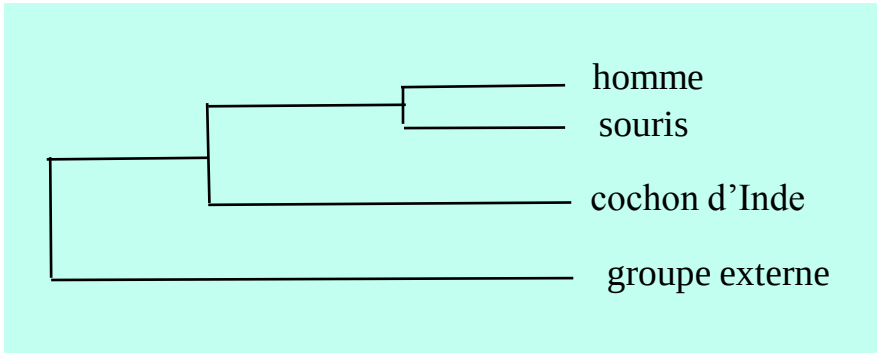
- Cet artéfact provient des inégalités du taux d'évolution des caractères entre les lignées analysées.

- Les espèces qui évoluent plus vite que les autres pour les caractères utilisés se traduisent dans un arbre par une branche propre plus longue.

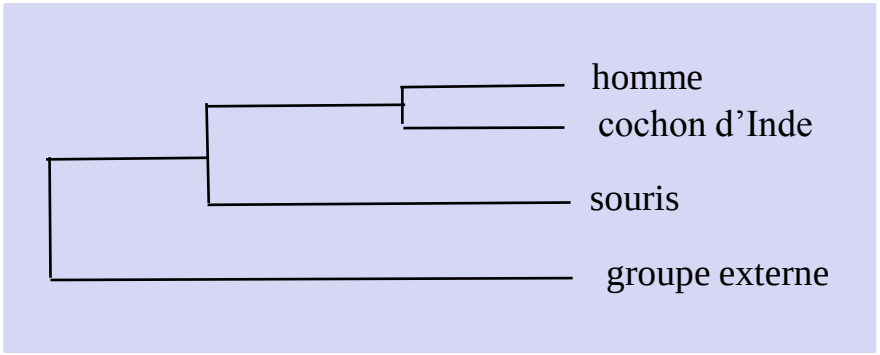
- On a pu montrer théoriquement et expérimentalement qu'au-delà d'un certain écart de vitesse d'évolution entre les espèces, les espèces qui évoluent plus vite ont plus de chance d'avoir des états de caractères communs que par ascendance commune, et que le nombre de caractères communs ainsi acquis devenait supérieur aux caractères qui auraient dû les séparer.

- Par conséquent, elles sont regroupées ensemble dans l'arbre indépendamment des parentés.

L'attraction des longues branches



1ère étude sur 15 gènes qui ont évolués plus vite chez le cochon d'Inde que chez la souris.



2ème étude sur l'ADN mitochondrial qui a évolué plus vite chez la souris que chez le cochon d'Inde.

Le groupe externe utilisé pour enraciner l'arbre présente une longue branche, ce qui en général le cas surtout si celui-ci est distant du point de vue évolutif. Dans ce cas, les longues branches du groupe d'étude sont attirées par celle du groupe externe.
Conseil : pour le groupe externe ne pas considérer une seule espèce mais plusieurs présentant des distances évolutives étalonnées pour « casser » les longues branches.

Phylogénomique : introduction

Les limites des phylogénies basées sur un seul gène

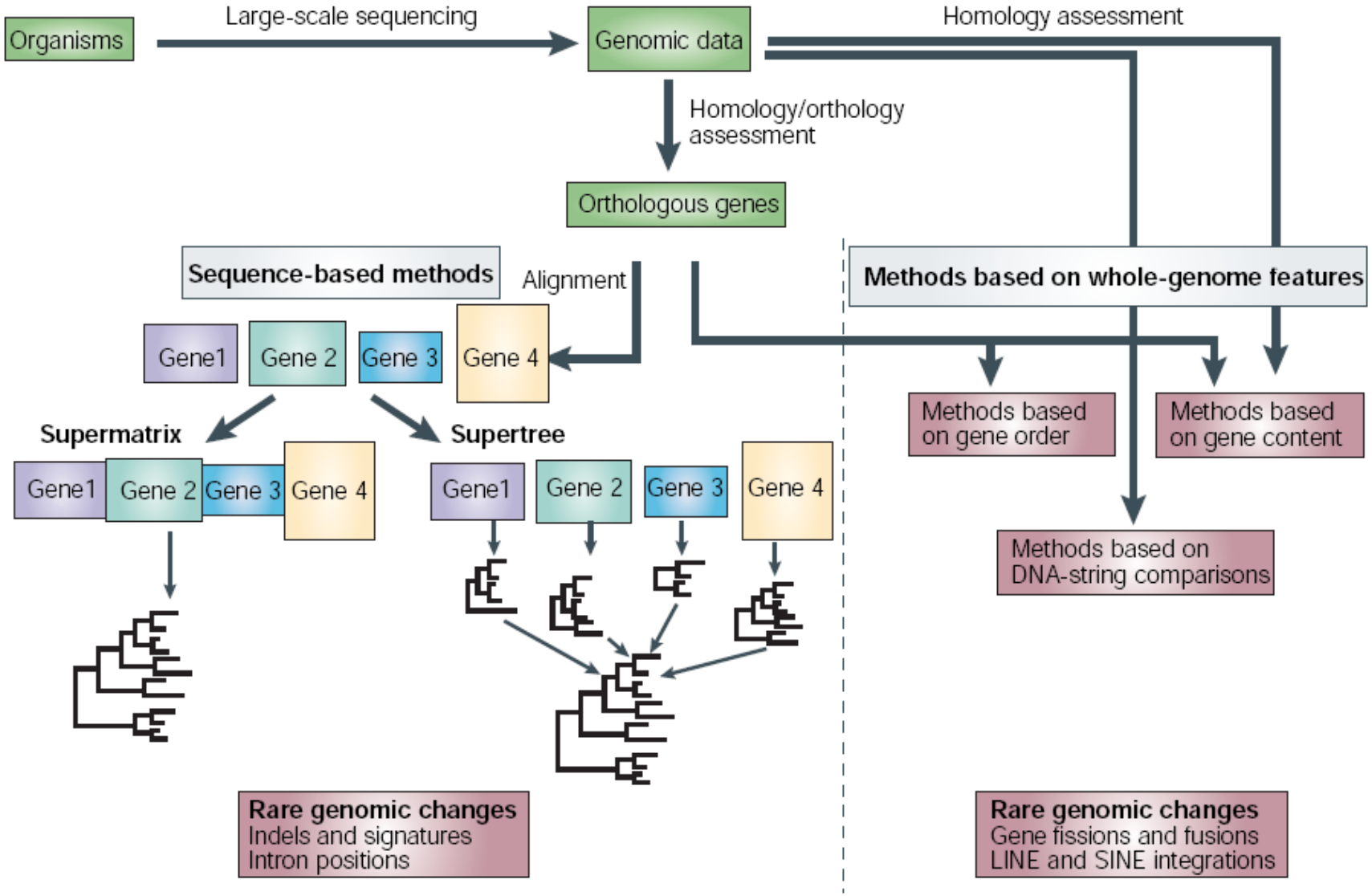
- Résolution limitée (erreur stochastique)

- 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

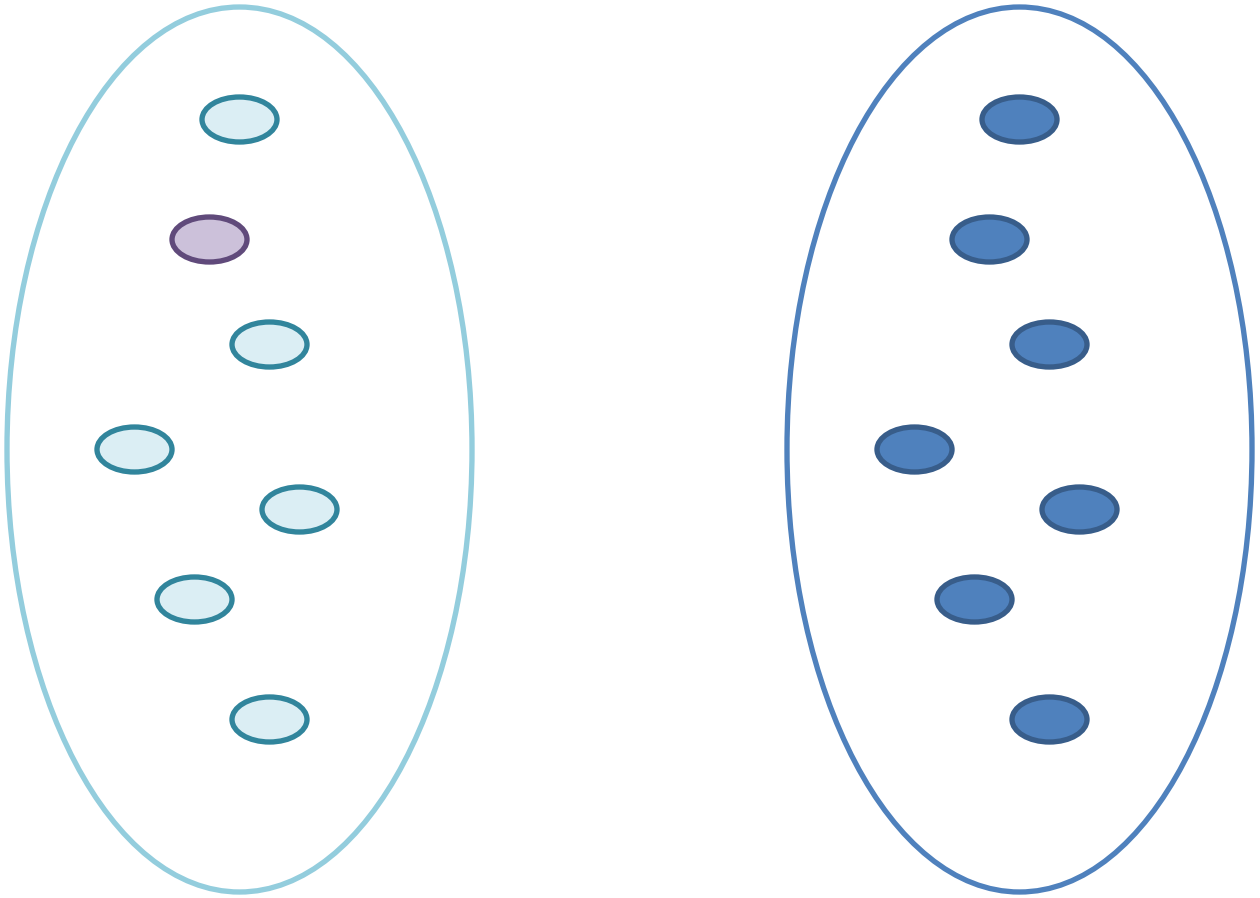
(Extrait de Delsuc *et al.*, 2005, *Nat. Rev. Genet. 6: 361-375)*



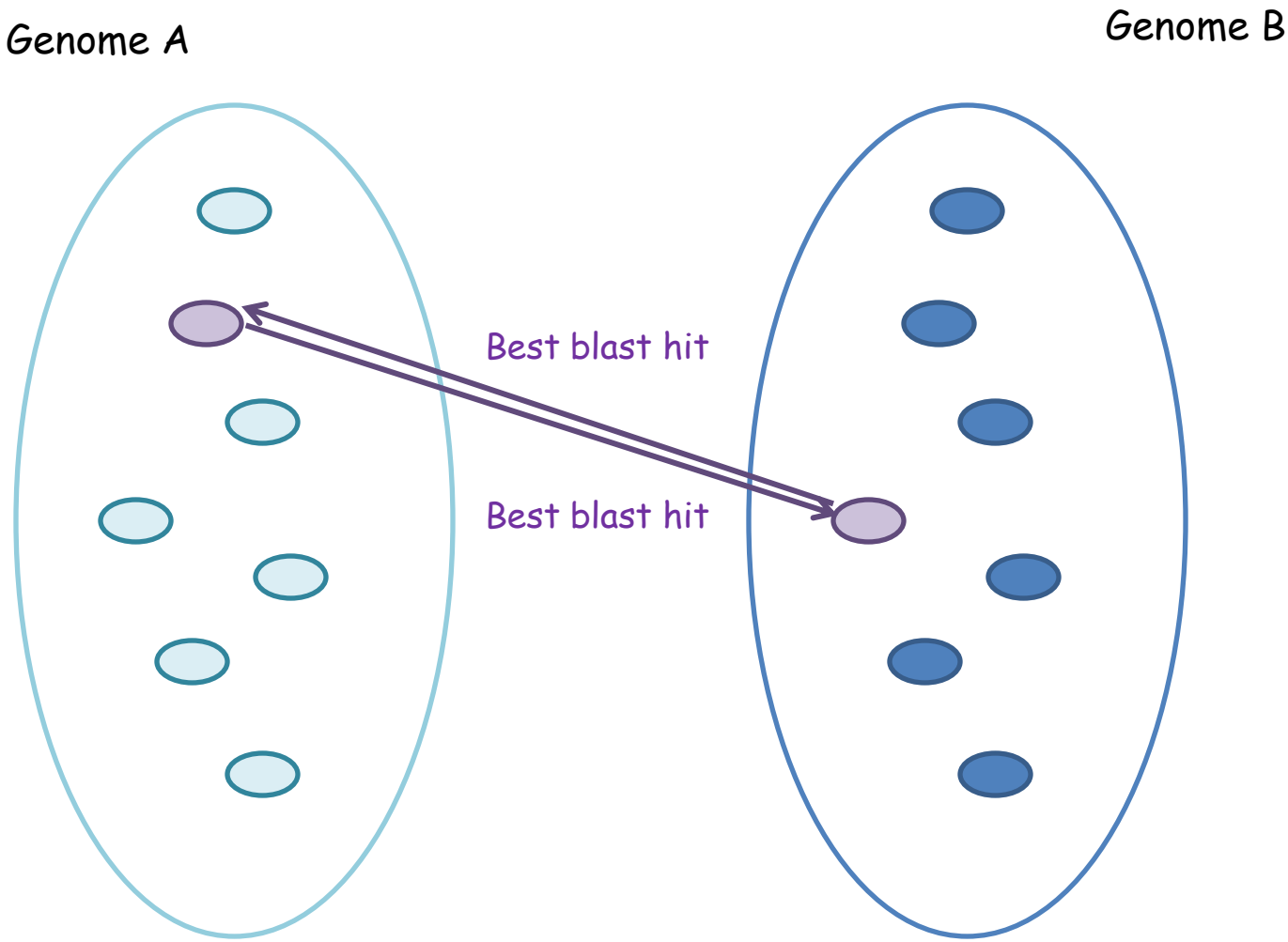
Orthologie en pratique

Genome A

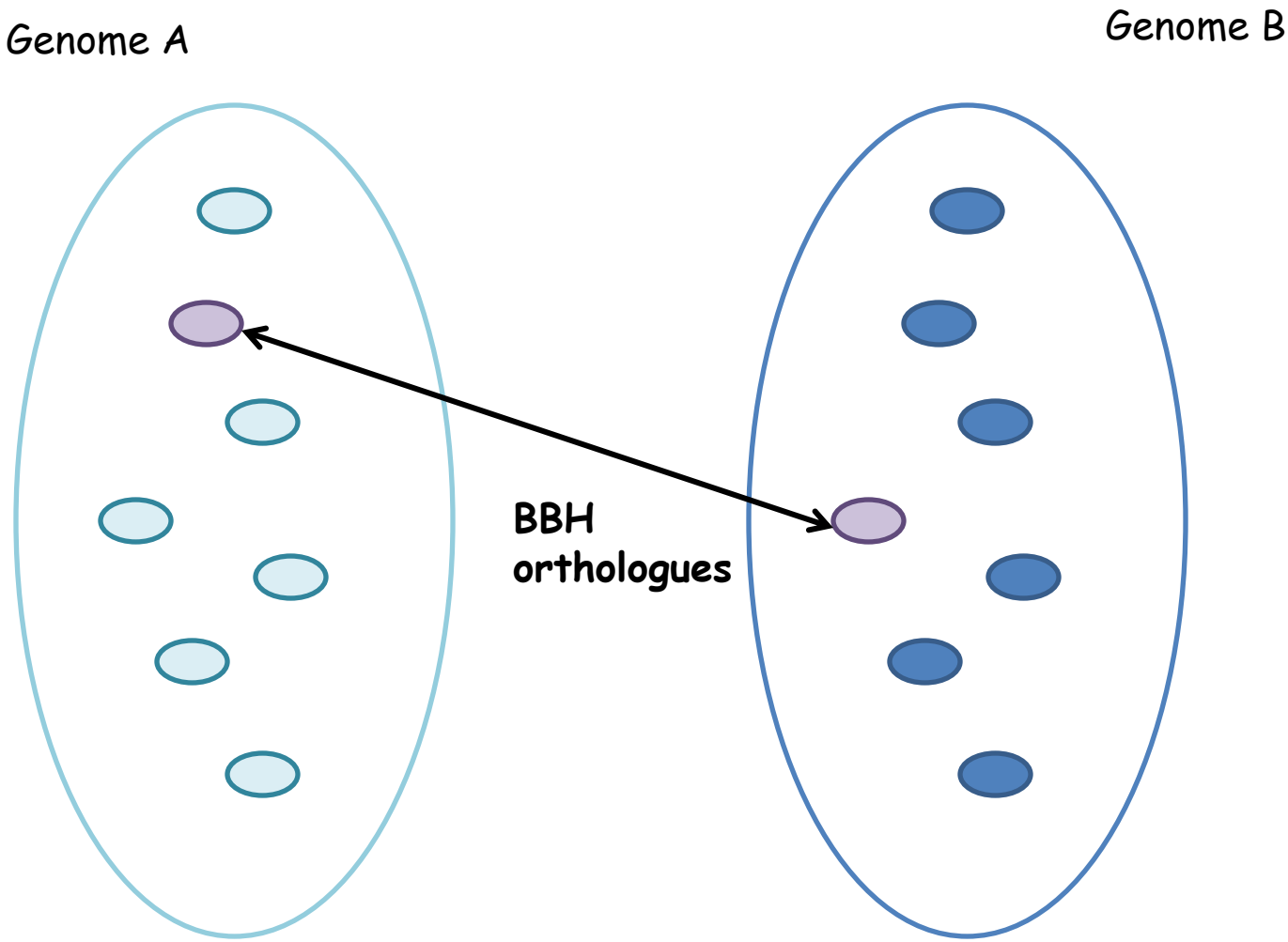
Genome B



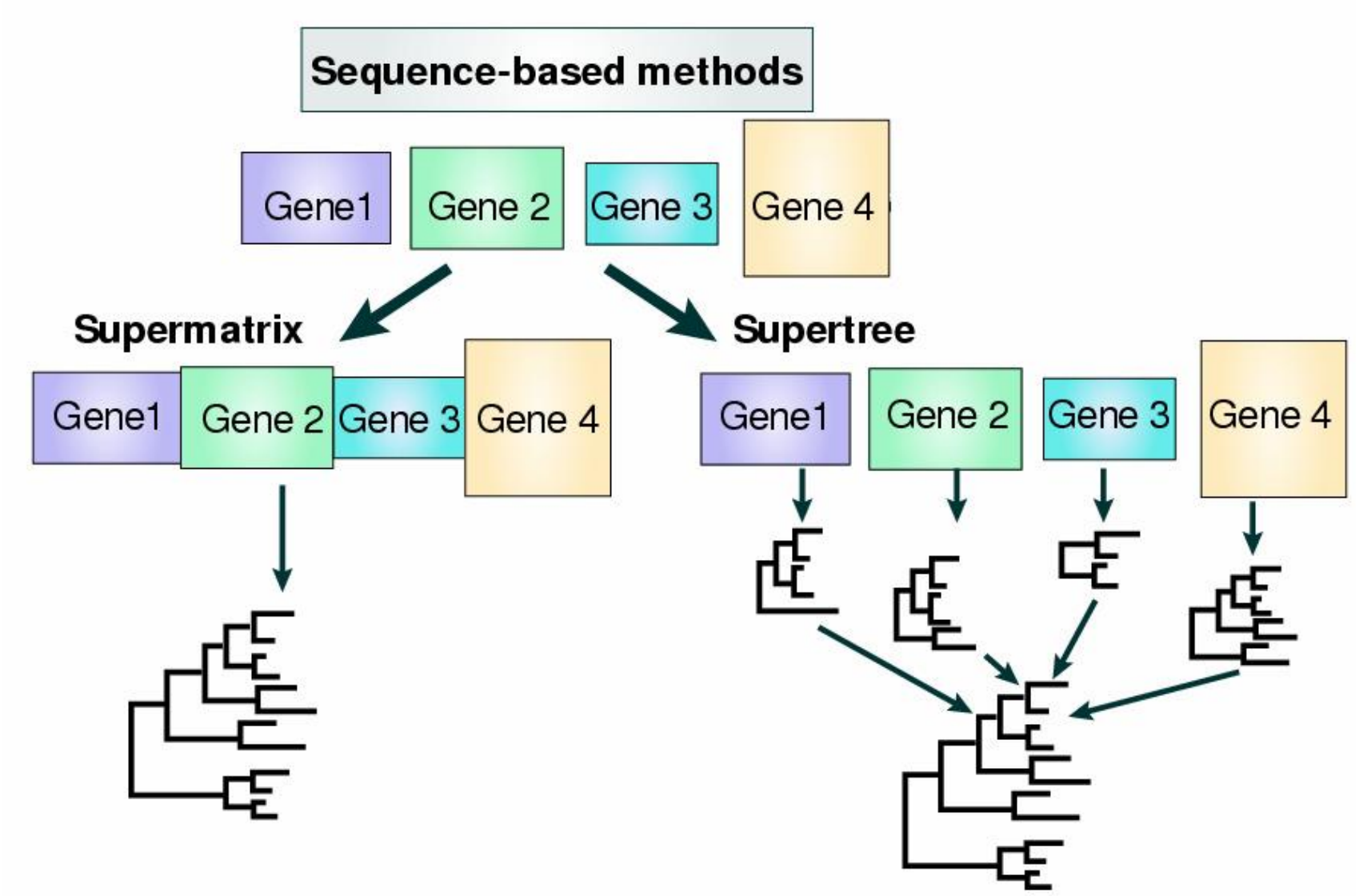
Orthologie en pratique



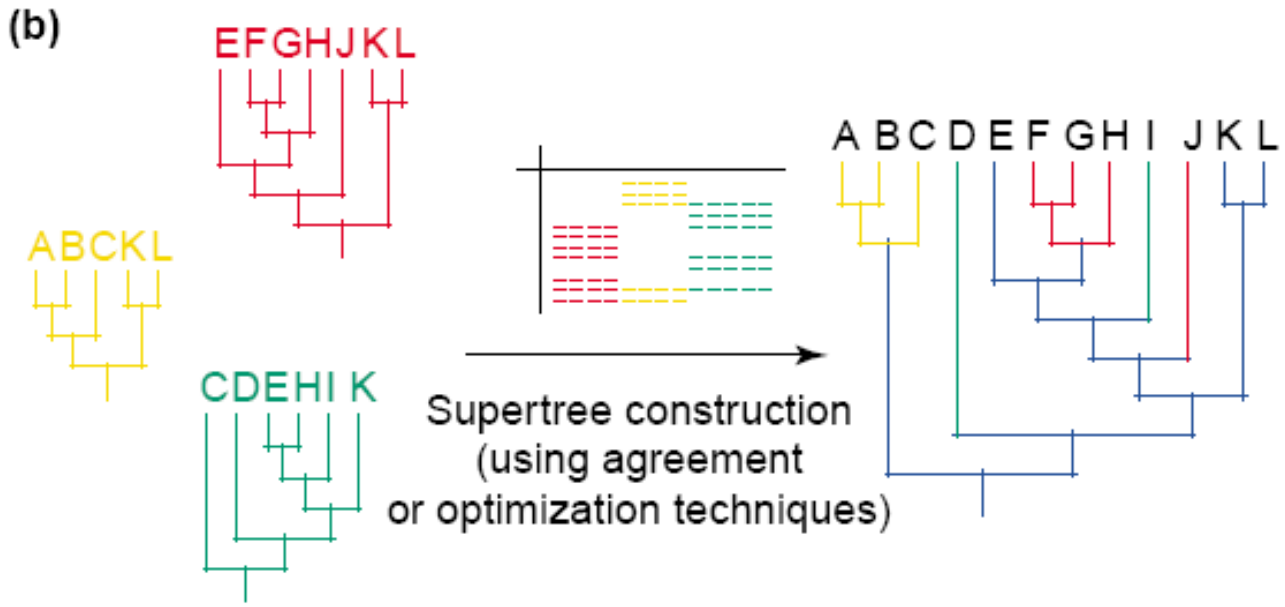
Orthologie en pratique



Méthodes basées sur les séquences : deux approches alternatives



Méthodes de superarbre



(Extrait de Bininda-Emonds, (2004), 19,315-322)

Principe :

- Construire des arbres individuels à partir de chaque matrice de caractères (alignements).
- Combiner les différents arbres définis sur des ensembles de taxa partiellement recouvrant, en un seul arbre, le superarbre.

Méthodes de superarbre

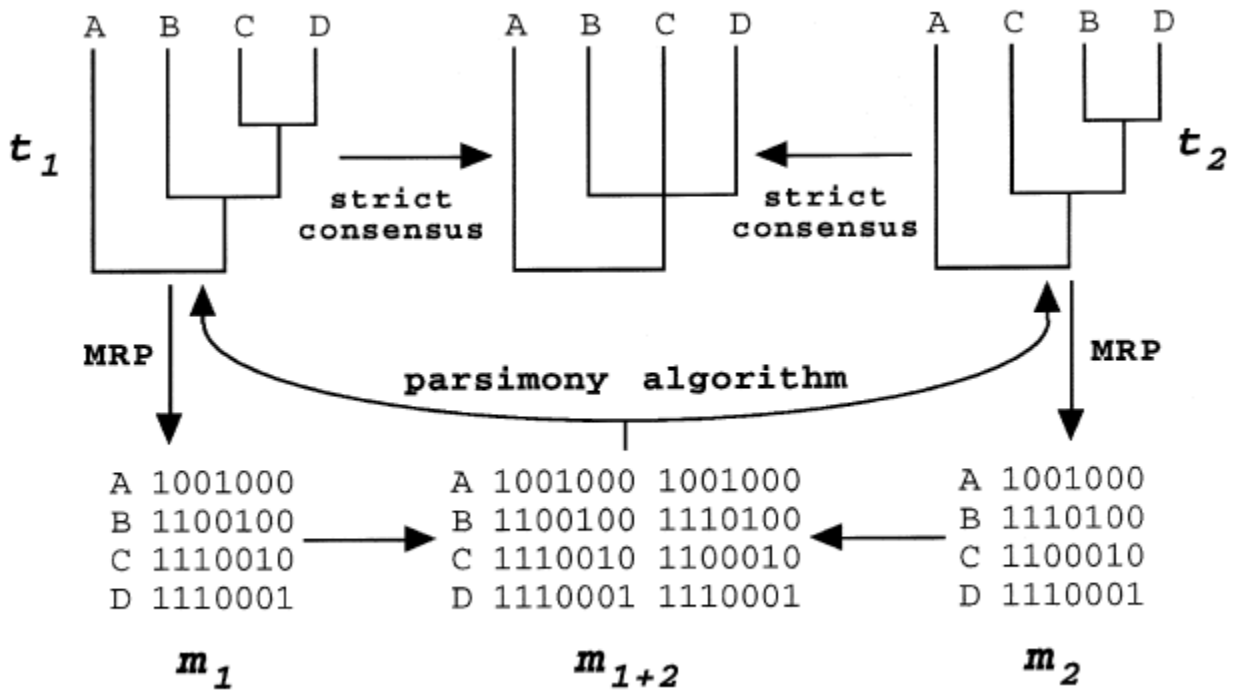
Il existe plusieurs méthodes de reconstruction des superarbres

Agreement supertrees	Refs	Optimization supertrees	Refs
MinCutSupertree	[50]	Average consensus (matrix representation using distances, MRD)	[51]
Modified mincut supertree	[52]	Bayesian supertrees	[46]
RankedTree	[53]	Gene tree parsimony	[36]
Semi-Labelled- and AncestralBuild	[15]	Matrix representation using compatibility (MRC)	[38,54]
Semi-strict	[25,55]	Matrix representation using flipping (MRF; also known as MinFlip supertrees)	[26]
Strict	[7]	Matrix representation using parsimony (MRP) and variants	[10,11,24,54,56]
Strict consensus merger	[47]	Most similar supertree method (dfit)	[a]
		Quartet supertrees	[28,57]

(Bininda-Emonds (2004). *Trends Ecol. Evol. 19: 315-322.)*

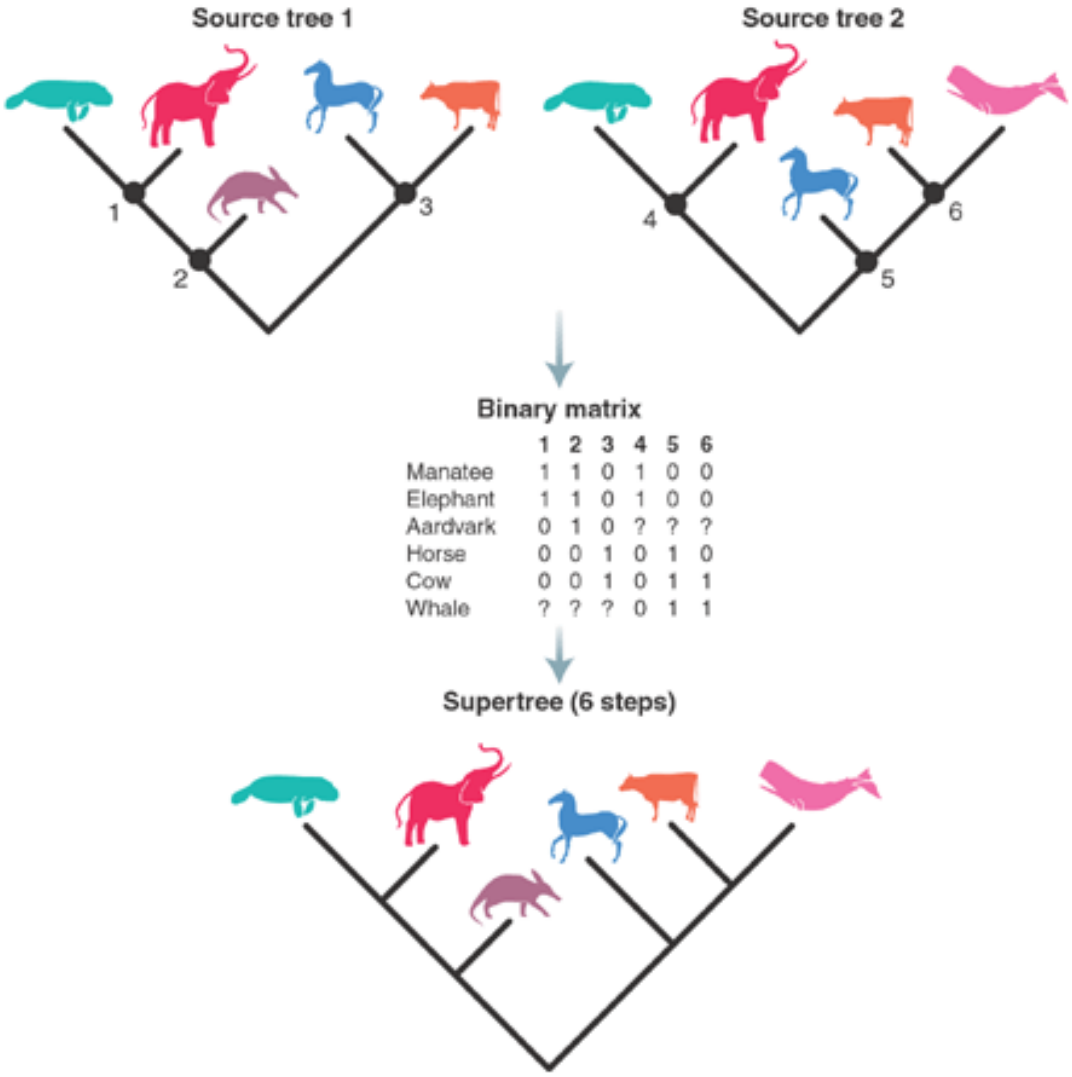
Cependant, la méthode Matrix Representation using Parsimony (MRP) est la plus utilisée, autant pour sa simplicité que son efficacité.

Méthodes de superarbre : Matrix Representation using Parsimony



Soit un arbre *t* représentant les relations entre un ensemble de taxa (feuilles) S={1,…,n}. Il est transformé en une matrice binaire à *n* lignes représentant les *n* feuilles et *p* colonnes, correspondant au nœuds internes.
Pour chaque nœud, les feuilles en descendant sont codées 1 et les autres 0. dans l'exemple, le codage des nœuds terminaux (feuilles) a été rajouté. On peut aussi pour polariser les éléments rajouter une ligne qu'avec des zéros (outgroup/racine) (méthode proposée par Baum (1992) et Ragan (1992))

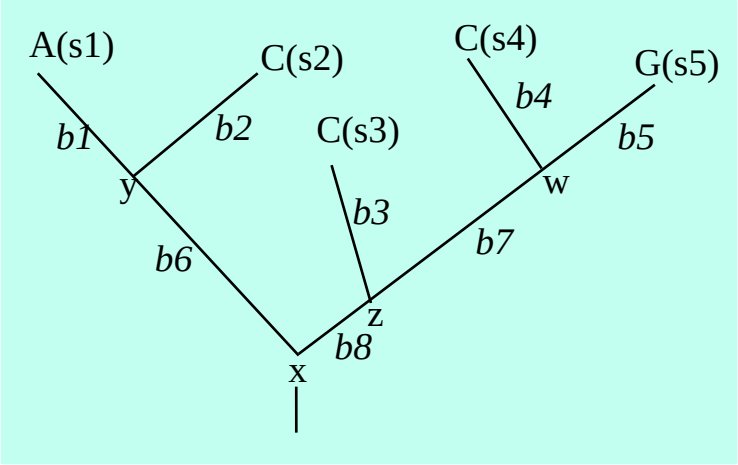
Méthodes de superarbre : Matrix Representation using Parsimony



Si certains taxa ne sont pas présents dans tous les arbres sources, ils sont représentés par des ? dans la matrice finale.

Maximum de vraisemblance : détails des calculs (pour les curieux)

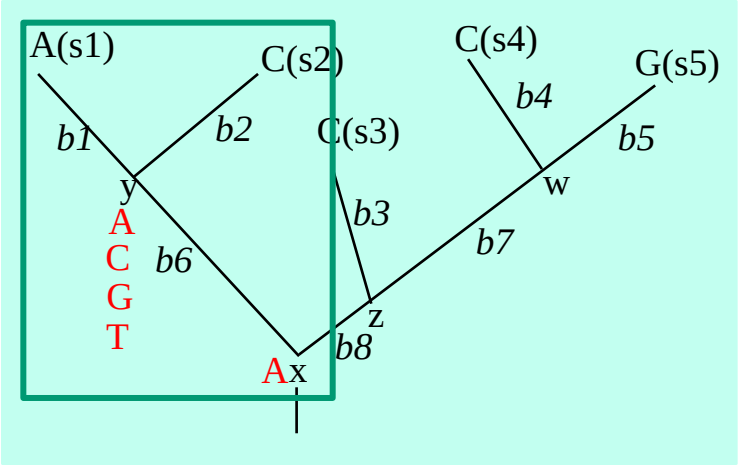
Maximum de vraisemblance



Connaissant le modèle évolutif, et partant de la racine, il est possible de calculer la probabilité d'occurrence de chaque base à chaque nœud interne et donc par extension celle de chaque scénario évolutif. On obtient donc la vraisemblance à un site en sommant l'ensemble des scénarios possibles

$$\mathrm{P}(D(i)\,|\,T) = \sum_{x} \mathrm{P}(x) \left(\sum_{y} P(y\,|\,x, b6) P(A\,|\,y, b1) P(C\,|\,y, b2) \times \left(\sum_{z} P(z\,|\,x, b8) P(C\,|\,z, b3) \times \left(\sum_{w} P(w\,|\,z, b7) P(C\,|\,w, b4) P(G\,|\,w, b5) \right) \right) \right)$$

Maximum de vraisemblance



Un scénario : x = A, y = A. La probabilité d'observer A(s1) et C(s2) sous ce scénario est :

$$P(A)\times\big(P(A\mid A,b6)P(A\mid A,b1)P(C\mid A,b2)\big)$$

autre scénario pour x = A, y = C

$$P(A)\times\big(P(C\mid A,b6)P(A\mid C,b1)P(C\mid C,b2)\big)$$

autre scénario pour x = A, y = G

$$P(A)\times\big(P(G\mid A,b6)P(A\mid G,b1)P(C\mid G,b2)\big)$$

autre scénario pour x = A, y = T

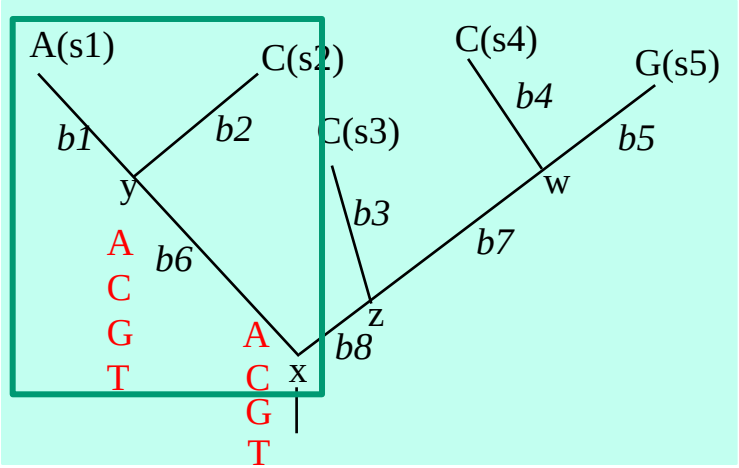
$$P(A)\times\big(P(T\mid A,b6)P(A\mid T,b1)P(C\mid T,b2)\big)$$

La probabilité d'observer A(s1) et C(s2) si x = A et y =A, C, G ou T est la somme des probabilités des 4 scénarios précédents soit :

$$P(A)\left(\sum_{y=A,C,G,T} P(y\mid A,b6)P(A\mid y,b1)P(C\mid y,b2)\right)$$

Maximum de vraisemblance

On a le même raisonnement si x = C, G ou T



La probabilité d'observer A(s1) et C(s2) si x = C et y = A, C, G ou T est :

$$P(C)\times\left(\sum_{y=A,C,G,T} P(y\,|\,C,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)$$

La probabilité d'observer A(s1) et C(s2) si x = G et y = A, C, G ou T est :

$$P(G)\times\left(\sum_{y=A,C,G,T} P(y\,|\,G,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)$$

La probabilité d'observer A(s1) et C(s2) si x = T et y = A, C, G ou T est :

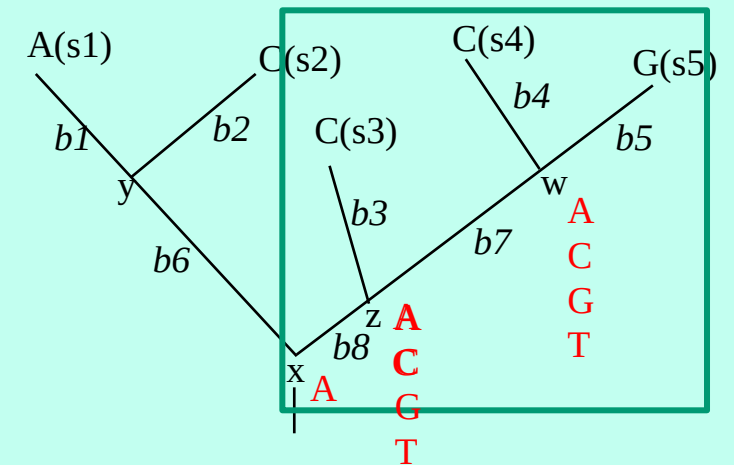
$$P(T)\times\left(\sum_{y=A,C,G,T} P(y\,|\,T,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)$$

La probabilité d'observer A(s1) et C(s2) en fonction des données au site i D(i) pour le sous-arbre est :

$$P(A)\left(\sum_{y=A,C,G,T} P(y\,|\,A,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)+P(C)\left(\sum_{y=A,C,G,T} P(y\,|\,A,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)+P(G)\left(\sum_{y=A,C,G,T} P(y\,|\,A,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)+P(T)\left(\sum_{y=A,C,G,T} P(y\,|\,A,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)$$

$$\sum_{X=A,C,G,T} P(x)\left(\sum_{y} P(y\,|\,x,b6)P(A\,|\,y,b1)P(C\,|\,y,b2)\right)$$

Maximum de vraisemblance



On va faire le même raisonnement sur l'autre sous-arbre mais un peu plus compliqué car un nœud interne supplémentaire.

La probabilité d'observer C(s3), C(s4) et G(s5) si x = A et z = A est :

$$P(A)\times\big(P(A|A,b8)\,P(C|A,b3)\times\begin{pmatrix}\big(P(A|A,b7)P(C|A,b4)P(G|A,b5)\big)\\+\big(P(C|A,b7)P(C|C,b4)P(G|C,b5)\big)\\+\big(P(G|A,b7)P(C|G,b4)P(G|G,b5)\big)\\+\big(P(T|A,b7)P(C|T,b4)P(G|T,b5)\big)\end{pmatrix}$$

Soit :

$$P(A)\times\big(P(A|A,b8)\,P(C|A,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)$$

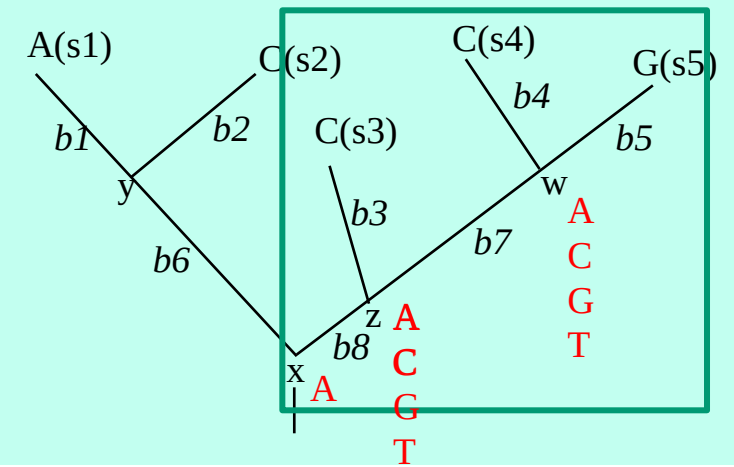
La probabilité d'observer C(s3), C(s4) et G(s5) si x = A et z = C est :

$$P(A)\times\big(P(C|A,b8)\,P(C|C,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)$$

On a la probabilité d'observer C(s3), C(s4), G(s5) si x = A et z =A, C, G ou T et w = A, C, G ou T

$$P(A)\times\big(P(A|A,b8)\,P(C|A,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)+P(A)\times\big(P(C|A,b8)\,P(C|C,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)+$$
$$P(A)\times\big(P(G|A,b8)\,P(C|G,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)+P(A)\times\big(P(T|A,b8)\,P(C|T,b3)\left(\sum_{w=A,C,G,T}P(w|z,b7)P(C|w,b4)P(G|w,b5)\right)$$

Maximum de vraisemblance



On va faire le même raisonnement sur l'autre sous-arbre mais un peu plus compliqué car un nœud interne supplémentaire.

La probabilité d'observer C(s3), C(s4) et G(s5) si x = A et z = A est :

$$P(A)\times\big(P(A\mid A,b8)\,P(C\mid A,b3)\times\begin{pmatrix}\big(P(A\mid A,b7)P(C\mid A,b4)P(G\mid A,b5)\big)\\+\big(P(C\mid A,b7)P(C\mid C,b4)P(G\mid C,b5)\big)\\+\big(P(G\mid A,b7)P(C\mid G,b4)P(G\mid G,b5)\big)\\+\big(P(T\mid A,b7)P(C\mid T,b4)P(G\mid T,b5)\big)\end{pmatrix}$$

Soit :

$$P(A)\times\big(P(A\mid A,b8)\,P(C\mid A,b3)\left(\sum_{w=A,C,G,T}P(w\mid z,b7)P(C\mid w,b4)P(G\mid w,b5)\right)$$

La probabilité d'observer C(s3), C(s4) et G(s5) si x = A et z = C est :

$$P(A)\times\big(P(C\mid A,b8)\,P(C\mid C,b3)\left(\sum_{w=A,C,G,T}P(w\mid z,b7)P(C\mid w,b4)P(G\mid w,b5)\right)$$

On a la probabilité d'observer C(s3), C(s4), G(s5) si x = A et z =A, C, G ou T et w = A, C, G ou T

$$P(A)\left(\sum_{Z=A,C,G,T}P(z\mid A,b8)P(C\mid z,b3)\left(\sum_{w=A,C,G,T}P(w\mid z,b7)P(C\mid w,b4)P(G\mid w,b5)\right)\right)$$

Maximum de vraisemblance

On a donc pour le sous-arbre gauche la probabilité d'observer A(s1) et C(s2) si x = A et y = A, C, G ou T :

$$P(A)\left(\sum_{y=A,C,G,T} P(y \mid A,b6)P(A \mid y,b1)P(C \mid y,b2)\right)$$

On a pour le sous arbre droit la probabilité d'observer C(s3), C(s4), G(s5) si x = A et z =A, C, G ou T et w = A, C, G ou T :

$$P(A)\left(\sum_{Z=A,C,G,T} P(z \mid A,b8)P(C \mid z,b3)\left(\sum_{w=A,C,G,T} P(w \mid z,b7)P(C \mid w,b4)P(G \mid w,b5)\right)\right)$$

La probabilité d'observer A(s1), C(s2), C(s3), C(s4), G(s5) si x = A et y =A, C, G ou T z =A, C, G ou T et w = A, C, G ou T est donné par le produit de ces deux probabilités, soit :

$$P(A)\left(\sum_{y=A,C,G,T} P(y \mid A,b6)P(A \mid y,b1)P(C \mid y,b2)\times\left(\sum_{Z=A,C,G,T} P(z \mid A,b8)P(C \mid z,b3)\left(\sum_{w=A,C,G,T} P(w \mid z,b7)P(C \mid w,b4)P(G \mid w,b5)\right)\right)\right)$$

La probabilité d'observer A(s1), C(s2), C(s3), C(s4) et G(s5) en fonction des données au site i D(i) est donc :

$$\mathrm{P}(D(i) \mid T) = \sum_x \mathrm{P}(x)\left(\sum_y P(y \mid x,b6)P(A \mid y,b1)P(C \mid y,b2)\times\left(\sum_z P(z \mid x,b8)P(C \mid z,b3)\times\left(\sum_w P(w \mid z,b7)P(C \mid w,b4)P(G \mid w,b5\right)\right)\right)$$

Equation (1)