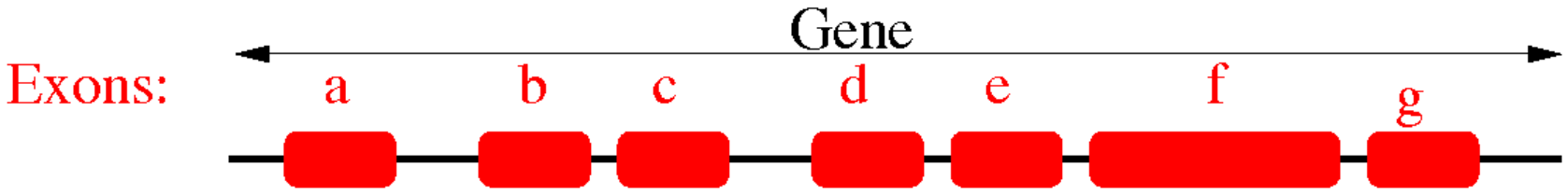
Les modèles phylogénétiques probabilistes et leurs applications

Samuel Blanquart, CR2 INRIA, EPI Bonsai

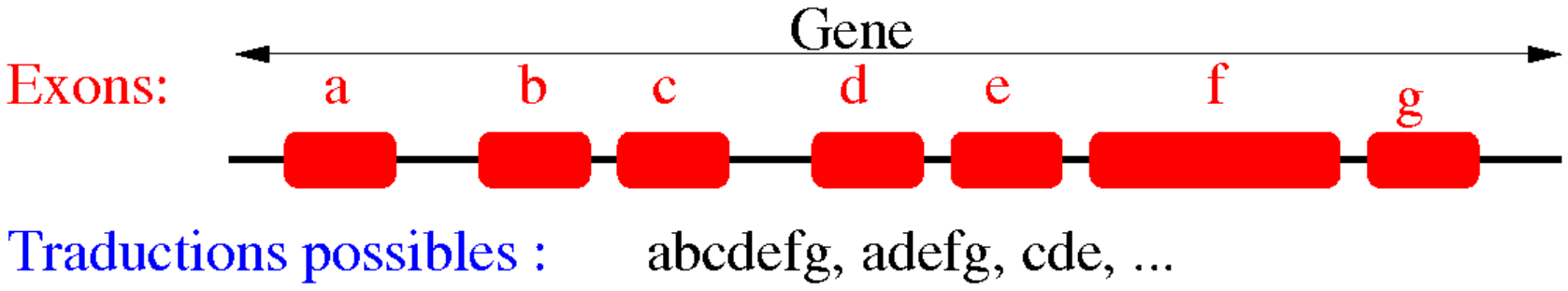
15 juin 2012

Journées du PPF Bioinformatique
Université de Lille 1

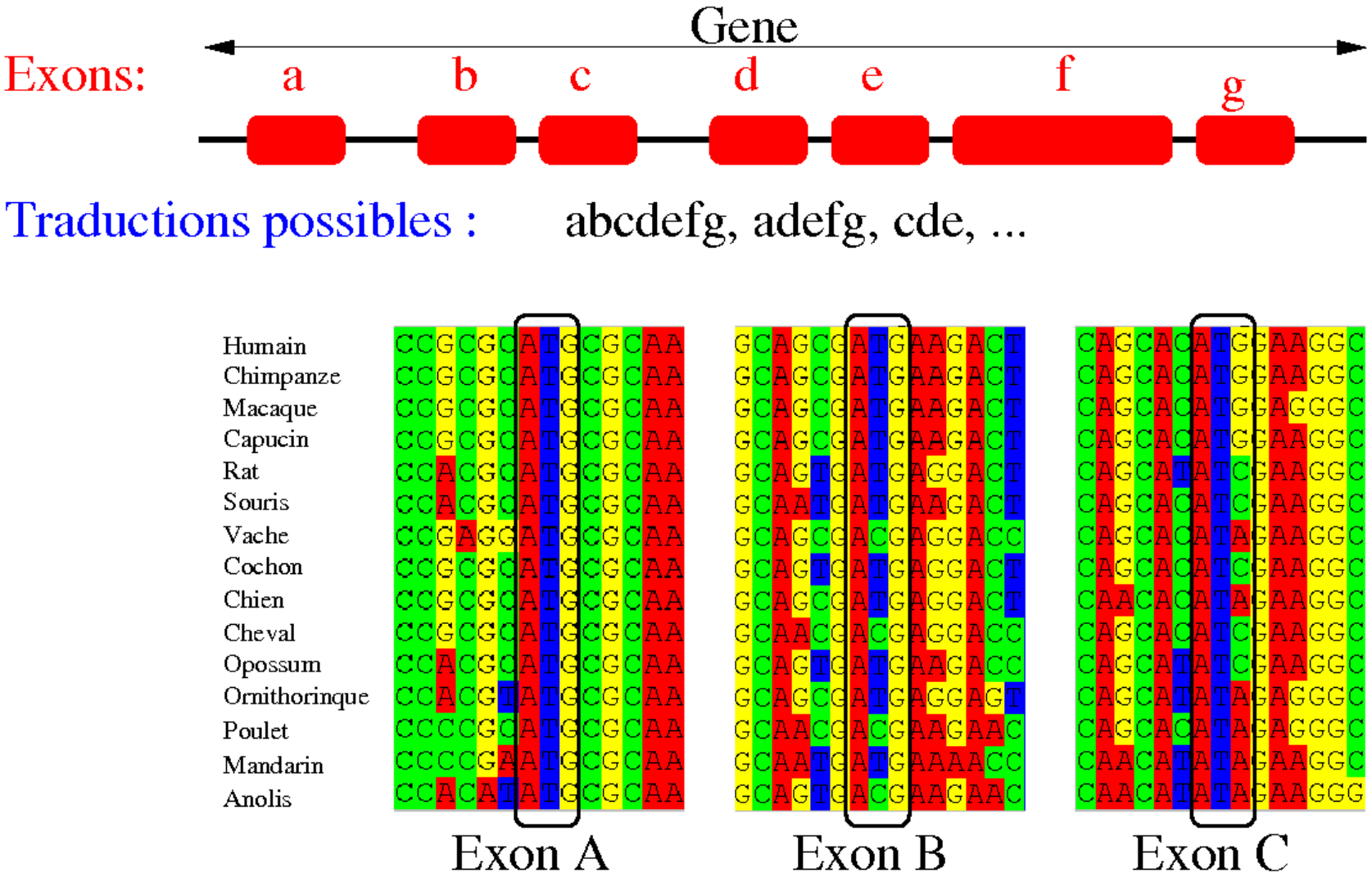
Divergence et homologie, cas des splicing alternatifs



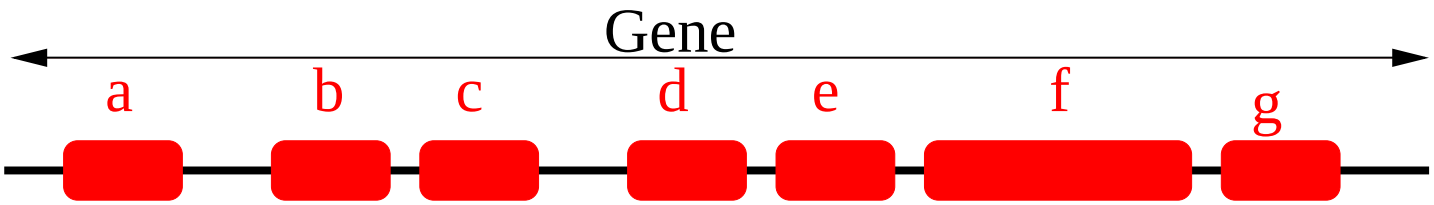
Divergence et homologie, cas des splicing alternatifs



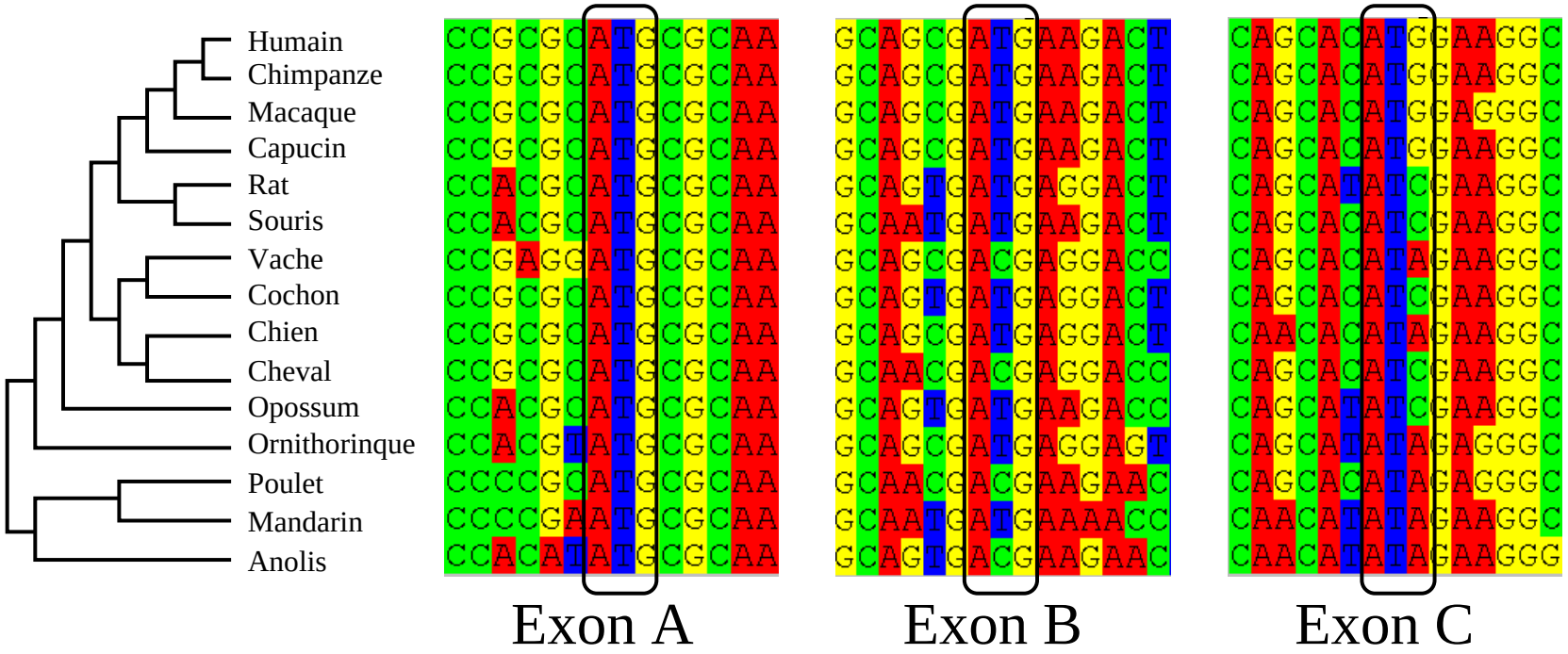
Divergence et homologie, cas des splicing alternatifs



Divergence et homologie, cas des splicing alternatifs



Traductions possibles : abcdefg, adefg, cde, ...



Caractères dérivés, caractères partagés

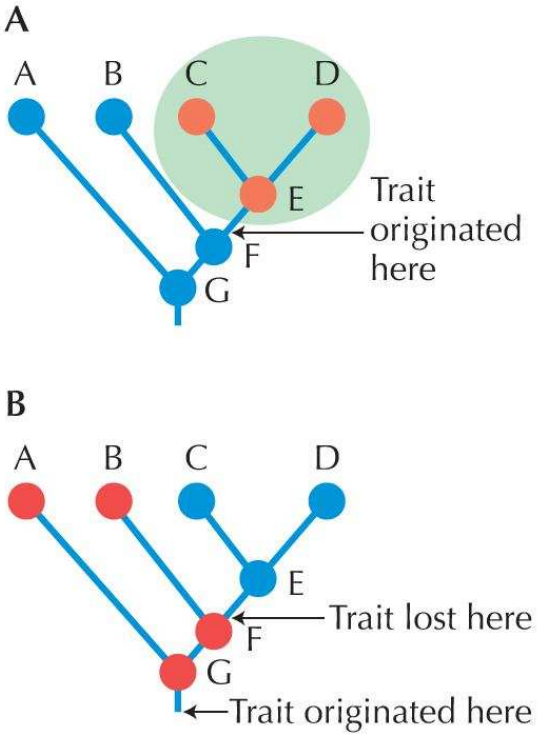


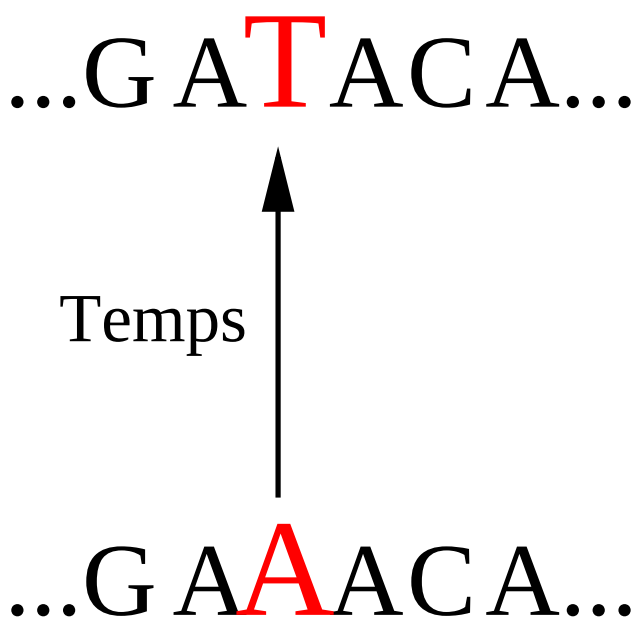
FIGURE 5.13. Ancestral and derived traits. (*A*) Derived traits. Consider a clade of species ABCD, with a trait present in species C and D but not A and B. If the trait evolved in the common ancestor of C and D (labeled E), then this is a derived trait in the group of ABCD because it was not present in the common ancestor of all four species. (*B*) Ancestral traits. Consider a clade of species ABCD, with a trait present in species A and B but not C and D. If the trait evolved in the common ancestor of A and B (labeled G), then this is an ancestral trait in the group of ABCD because it was present in the common ancestor of all four species. (Incidentally, the trait was subsequently lost in the lineage ECD.)

Evolution © 2007 Cold Spring Harbor Laboratory Press

Partie 1 : Les processus stochastiques en modélisation phylogénétique

Des processus de Markov comme modèles de l'évolution

Soit l'alphabet ADN, ayant 4 états : $\{A, C, G, T\}$.



Soit un processus Markovien Q :

$\rightarrow$	A	C	G	T
A	*	a	b	c
C	d	*	e	f
G	g	h	*	i
T	j	k	l	*

$Q_{i \rightarrow j}$ spécifie le taux instantané des substitutions $i \rightarrow j$

- * définition des cellules diagonales : $Q_{i \rightarrow i} = -\sum_{j \neq i} Q_{i \rightarrow j}$,
- Probabilité d'une substitution en un temps t :

$$P(i \rightarrow j | t) = [e^{t \times Q}]_{i,j}$$

Un processus de substitution sans paramètre : JC69

- Jukes et Cantor (1969),
- Hypothèse : toutes les substitutions surviennent selon le même taux,
- Propriété : les compositions nucléotidiques attendues sont uniformes : $\pi = \{0.25, 0.25, 0.25, 0.25\}$,
- 0 paramètre libre.

$$Q = \begin{pmatrix} * & 0.25 & 0.25 & 0.25 \\ 0.25 & * & 0.25 & 0.25 \\ 0.25 & 0.25 & * & 0.25 \\ 0.25 & 0.25 & 0.25 & * \end{pmatrix}$$

Modèle F81, des fréquences nucléotidiques variables

- Felsenstein (1981),
- Hypothèse : généralement, on n'observe pas les nucléotides A, C, G et T avec des fréquences égales. Par exemple, certains génomes, certaines régions génomiques, sont plus ou moins riches en $A + T$.
- Propriété : des paramètres $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ estiment les fréquences d'équilibre du processus de Markov (ou encore ses probabilités stationnaires), en fonction des compositions nucléotidiques des données.
- 3 paramètres libres ($\pi_T = 1 - \pi_A - \pi_C - \pi_G$).
- Si $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, F81 se réduit à JC69.

$$Q = \begin{pmatrix} * & \pi_C & \pi_G & \pi_T \\ \pi_A & * & \pi_G & \pi_T \\ \pi_A & \pi_C & * & \pi_T \\ \pi_A & \pi_C & \pi_G & * \end{pmatrix}$$

Modèle HKY85, des “transitions” plus fréquentes

- Hasegawa, Kishino et Yano (1985),
- Hypothèse : les transitions $A \leftrightarrow G$ (entre purines R) et $C \leftrightarrow T$ (entre pyrimidines Y) sont plus fréquentes que les transversions $R \leftrightarrow Y$,
- Propriétés : un paramètre additionnel κ estime un ratio transition/transversion, en fonction des données. Les compositions attendues sont spécifiées par les fréquences d'équilibre π.
- 4 paramètres libres : 3 ($\pi_T = 1 - \pi_A - \pi_C - \pi_G$) + 1 ($\kappa$).
- Si $\kappa = 1$, HKY85 se réduit à F81.

$$Q = \begin{pmatrix} * & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & * \end{pmatrix}$$

Modèle GTR, “General Time Reversible”

- Lanave, Preparata, Saccone et Serio (1984).
- Hypothèse : les échangeabilités entre paires d'états sont plus générales que supposé par HKY85.
- Propriété : 6 paramètres $\rho_{i \leftrightarrow j}$ spécifient les échangeabilités entre les 12 paires de nucléotides (on suppose que $\rho_{i \to j} = \rho_{j \to i}$, le processus est dit réversible).
- 8 paramètres libres : 3 $(\pi_T = 1 - \pi_A - \pi_C - \pi_G) + 5$ $(\rho_{GT} = 1 - \rho_{AC} - \rho_{AG} - \rho_{AT} - \rho_{CG} - \rho_{CT})$.
- Si $\rho_{AG} = \rho_{CT} = \kappa$ et $\rho_{GT} = \rho_{AC} = \rho_{AT} = \rho_{CG} = 1$, GTR se réduit à HKY85.

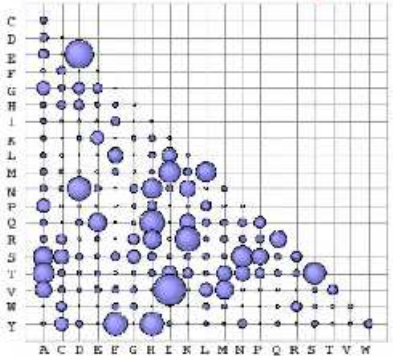
$$Q = \begin{pmatrix} * & \rho_{AC}.\pi_C & \rho_{AG}.\pi_G & \rho_{AT}.\pi_T \\ \rho_{AC}.\pi_A & * & \rho_{CG}.\pi_G & \rho_{CT}.\pi_T \\ \rho_{AG}.\pi_A & \rho_{CG}.\pi_C & * & \rho_{GT}.\pi_T \\ \rho_{AT}.\pi_A & \rho_{CT}.\pi_C & \rho_{GT}.\pi_G & * \end{pmatrix}$$

Modèle d'évolution des séquences protéiques

- 20 acides aminés,
- Le processus Markovien Q est une matrice 20×20,
- Les 20 fréquences d'équilibre du processus sont spécifiées par un vecteur π.
- Les échangeabilités entre chaque paires d'acides aminés sont spécifiées par une matrice ρ (symétrique $\rho_{i \to j} = \rho_{j \to i}$).
- 208 paramètres libres pour le modèle GTR : 19 (π) + 189 (ρ).

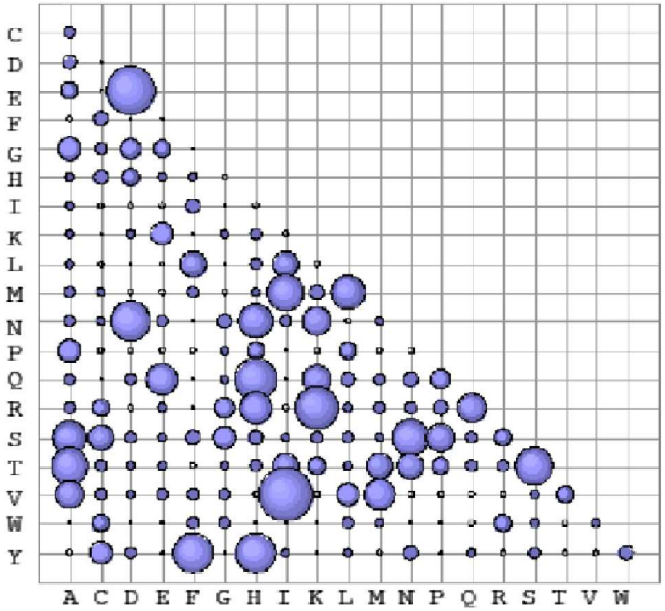
Probabilités Stationnaires π

Taux d'échanges relatifs ρ

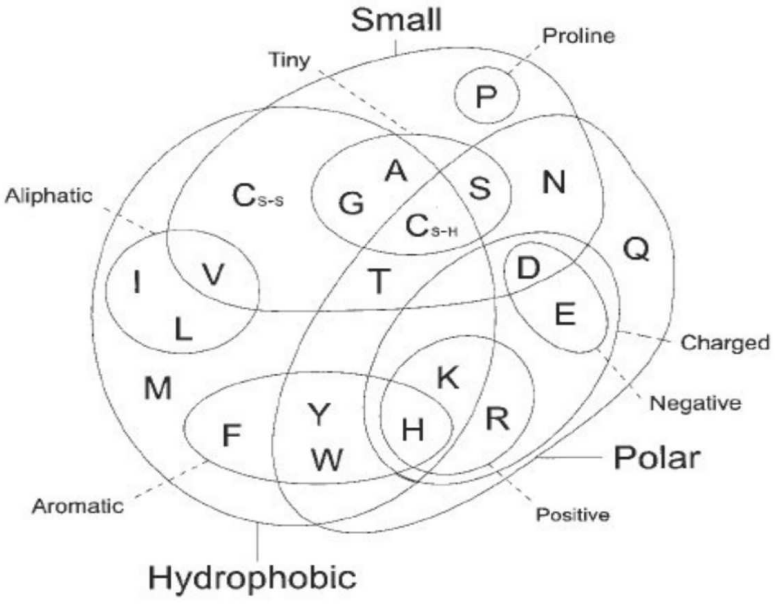


$$\begin{cases} Q_{l \neq m} = \pi_m \rho_{lm} \\ Q_{ll} = -\sum_{m \neq l} Q_{lm} \end{cases}$$

Echangeabilités et propriétés biochimiques



A



B

A) Modèle JTT, B) Diagramme de Venn des propriétés biochimiques (Taylor 1986). Les échanges entre acides aminés biochimiquement similaires sont plus fréquents (ex. I, V, L, M, F, Y hydrophobes, D, E chargés négativement, K, R chargés positivement, etc.).

Les matrices empiriques de remplacement

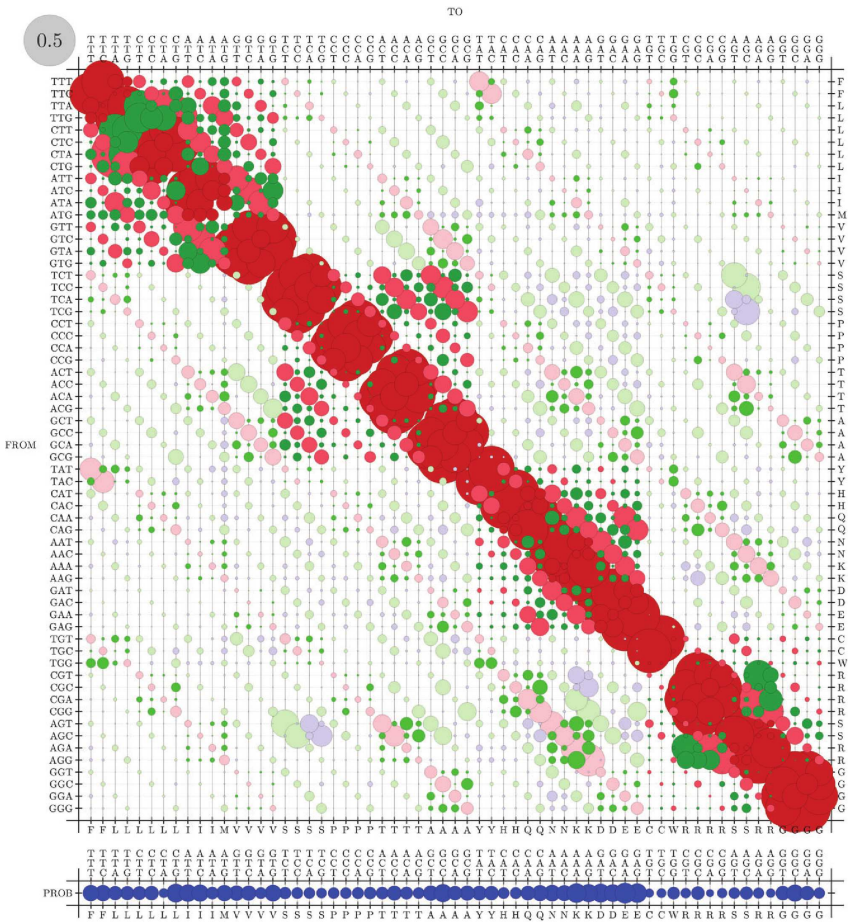
- Les vieilles à antiques : JTT, WAG, Blosum62, Dayhoff,
- L'universelle : LG,
- Les mitochondriales : MtREV (vertébrés), MtMam (mammifères), MtArt (arthropodes),
- La chloroplastique : CpREV,
- Pour les virus du SIDA : HIVb, HIVw,
- Pour les virus de la grippe : FLU,
- Etc.

Modèles à codons, le code génétique

	U		C		A		G		
U	UUU UUC UUA UUG	phénylalanine leucine	UCU UCC UCA UCG	sérine	UAU UAC UAA UAG	tyrosine stop	UGU UGC UGA UGG	cystéine stop tryptophane	U C A G
C	CUU CUC CUA CUG	leucine	CCU CCC CCA CCG	proline	CAU CAC CAA CAG	histidine glutamine	CGU CGC CGA CGG	arginine	U C A G
A	AUU AUC AUA AUG	isoleucine méthionine/start	ACU ACC ACA ACG	thréonine	AAU AAC AAA AAG	asparagine lysine	AGU AGC AGA AGG	sérine arginine	U C A G
G	GUU GUC GUA GUG	valine	GCU GCC GCA GCG	alanine	GAU GAC GAA GAG	acide aspartique acide glutamique	GGU GGC GGA GGG	glycine	U C A G

Code génétique universel. $4^3 = 64$ triplets de nucléotides (moins 3 triplets "stop"), ou codons , codent pour chacun des 20 acides aminés.

Modèles à codons empiriques (Kosiol et al. 2007)



Echangeabilités empiriques entre codons (1829 paramètres).
Substitutions simples (ex. $TTT \leftrightarrow TTA$, position 3, position 1, position 2), doubles (ex. $TTT \leftrightarrow TAA$) et triples (ex. $TTT \leftrightarrow AAA$).

Modèles à codons mécanistes (Goldman and Yang 1994 ; Nielsen and Yang 2003)

$$Q_{l \neq m} = \begin{cases} 0 & \text{si 2 ou 3 paires } (i_1, j_1),\ (i_2, j_2),\ (i_3, j_3) \text{ sont différentes,} \\ \pi_m & \text{si la substitution est synonyme et si c'est une transversion,} \\ \pi_m \kappa & \text{si la substitution est synonyme et si c'est une transition,} \\ \pi_m \omega & \text{si la substitution est non synonyme et si c'est une transversion,} \\ \pi_m \omega \kappa & \text{si la substitution est non synonyme et si c'est une transition.} \end{cases}$$

$$Q_{ll} = -\sum_{l \neq m} Q_{l,m}$$

Où :

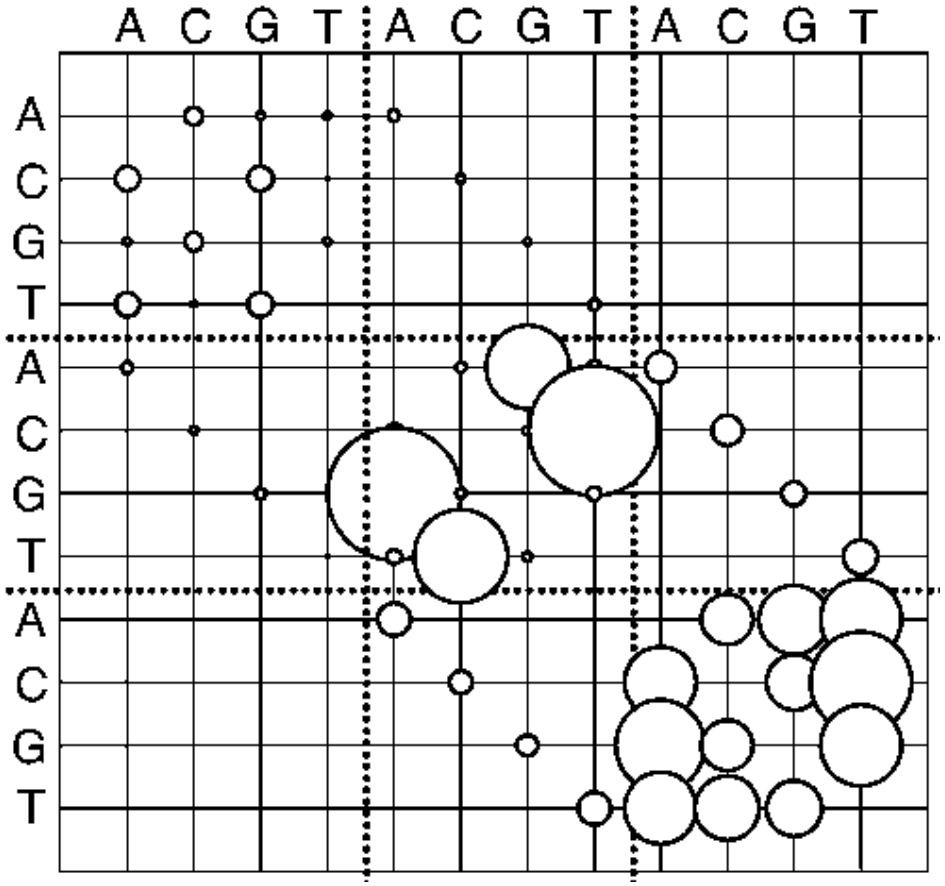
- π_m est la fréquence d'équilibre du mième codon ($m \in [1..61]$),
- κ est le taux de transition (taux fixé à 1 pour les transversions)
- ω est le taux de mutation non synonyme (taux fixé à 1 pour les mutations synonymes)

Modèles à codons mécanistes (Goldman and Yang 1994 ; Nielsen and Yang 2003)

$\omega = \frac{dN}{dS}$, où dN et dS sont respectivement les taux de mutations non-synonymes et synonymes, est un indicateur du régime de sélection naturelle.

- $\omega << 1$ indique une sélection négative tendant à conserver les phénotypes protéiques (peu de mutation non-synonymes),
- $\omega \approx 1$ indique un régime de sélection neutre (autant de mutation synonymes et non-synonymes),
- $\omega >> 1$ indique une pression de sélection positive tendant à favoriser les nouveaux phénotypes protéiques (excès de mutations non-synonymes).

Processus de substitution Markov Markov-Modulé

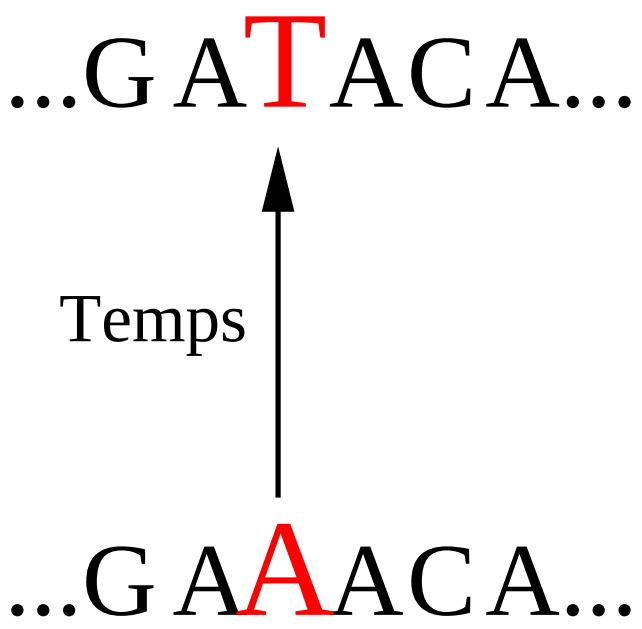


(Whelan 2008) Modèle MMM à 3 processus, appliqué à des régions codantes. Q_1 : lent et général (position 2), Q_2 transitions rapides, Q_3 rapide et général.

Partie 2 : Le modèle probabiliste standard et le calcul de la vraisemblance

Des processus de Markov comme modèles de l'évolution

Soit l'alphabet ADN, ayant 4 états : $\{A, C, G, T\}$.



Soit un processus Markovien Q :

$\rightarrow$	A	C	G	T
A	*	a	b	c
C	d	*	e	f
G	g	h	*	i
T	j	k	l	*

$Q_{i \rightarrow j}$ spécifie le taux instantané des substitutions $i \rightarrow j$

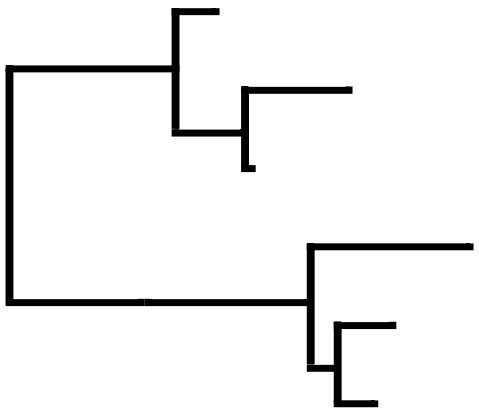
- * définition des cellules diagonales : $Q_{i \rightarrow i} = -\sum_{j \neq i} Q_{i \rightarrow j}$,
- Probabilité d'une substitution en un temps t :

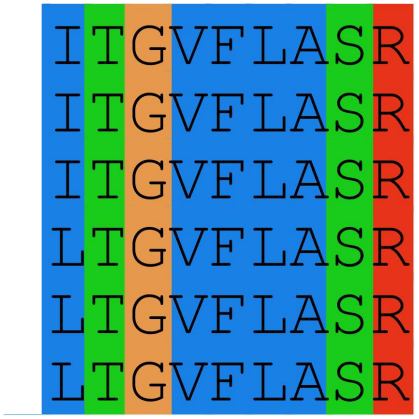
$$P(i \rightarrow j | t) = [e^{t \times Q}]_{i,j}$$

Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$





Données D

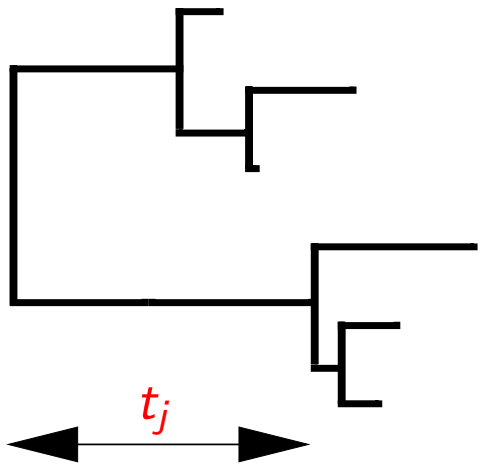
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ

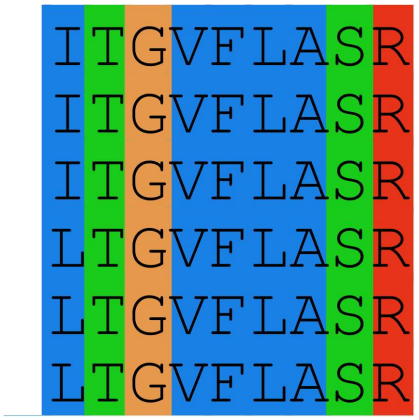
Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$

$\mathbf{t},$





Données D

- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ, Vitesses d'évolution $\mathbf{t}$,

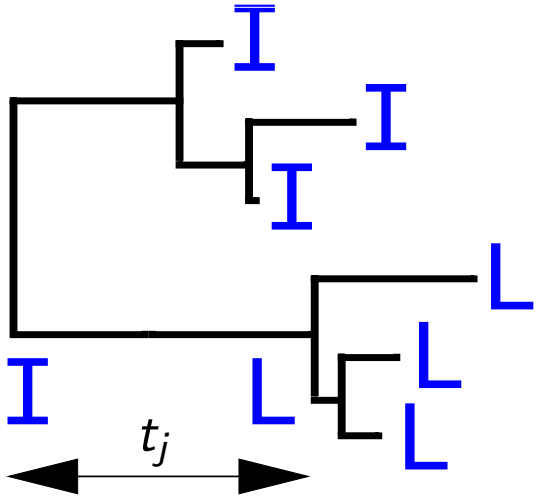
Le modèle phylogénétique standard

Modèle θ

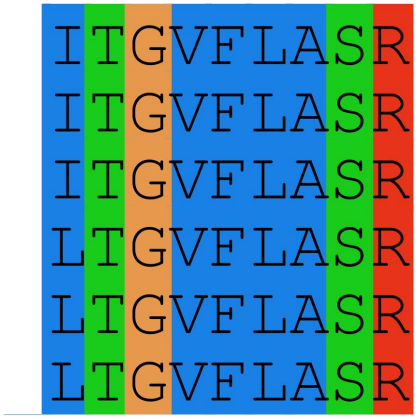
$\theta = \{\tau,$

$\mathbf{t},$

$Q\ \}$



$p(I \to L \mid Q, t_j)$



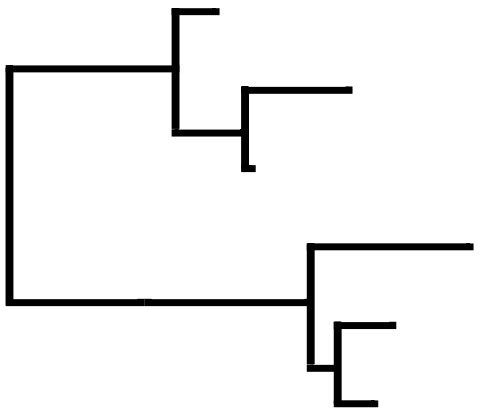
Données D

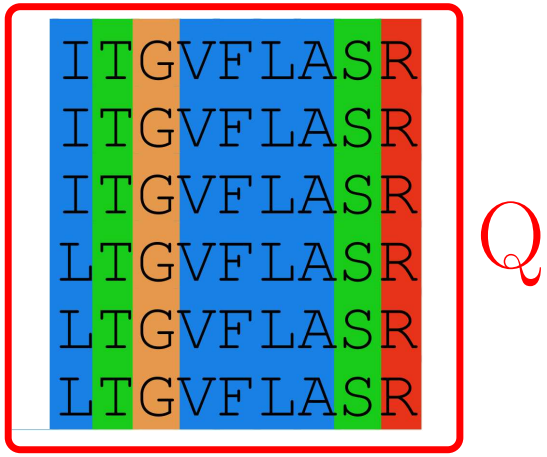
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ, Vitesses d'évolution $\mathbf{t}$,
- Q, générateur Markovien du processus de substitution.

Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau, \mathbf{t}, Q\}$

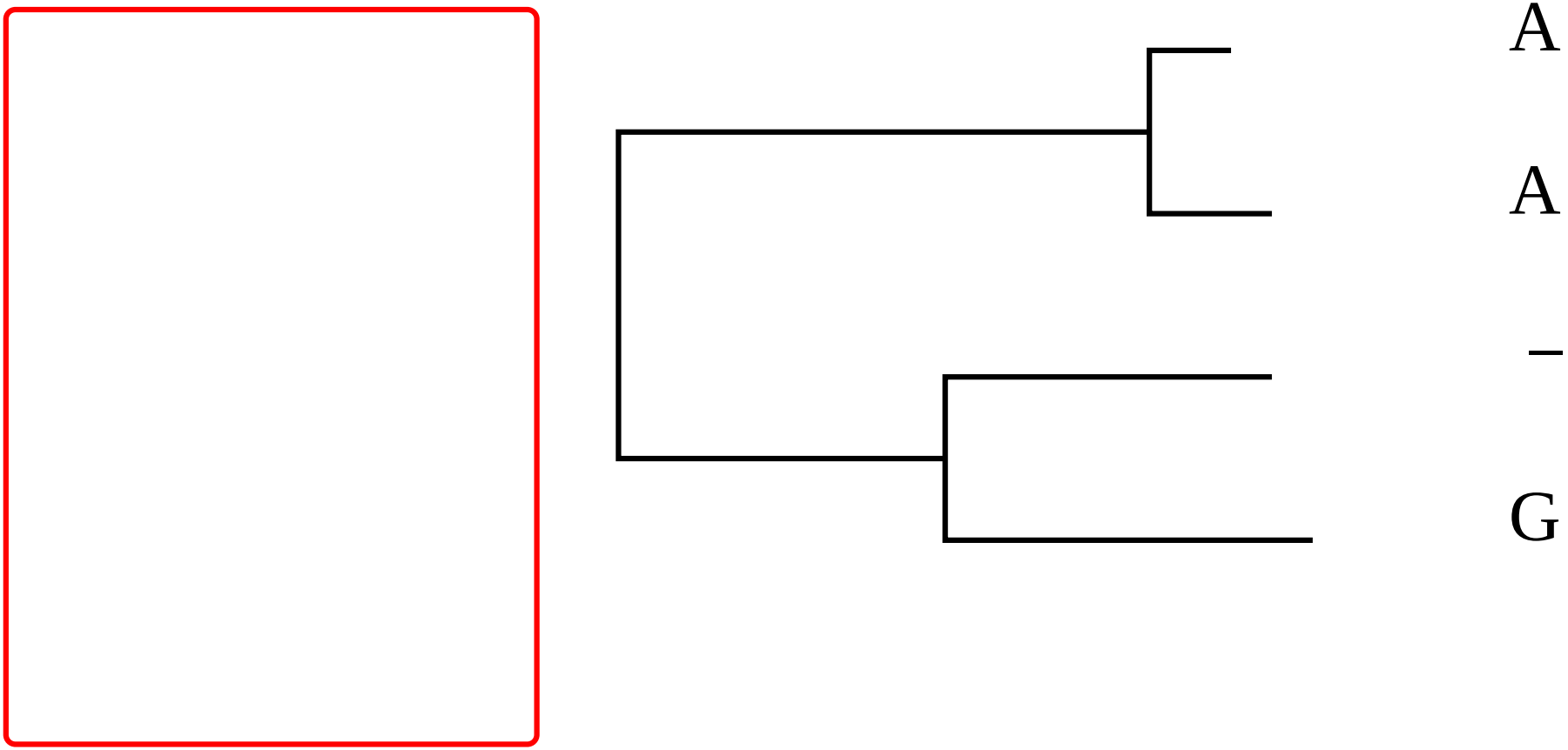




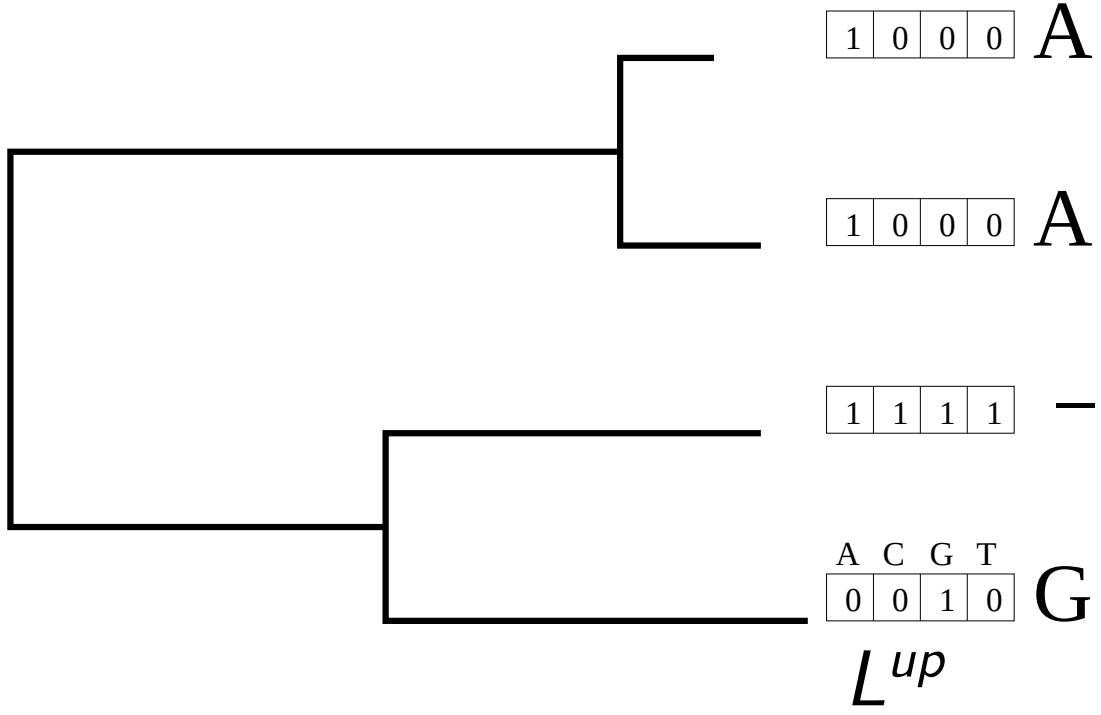
Données D

- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ, Vitesses d'évolution $\mathbf{t}$,
- Q, UNIQUE générateur Markovien du processus de substitution.

Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

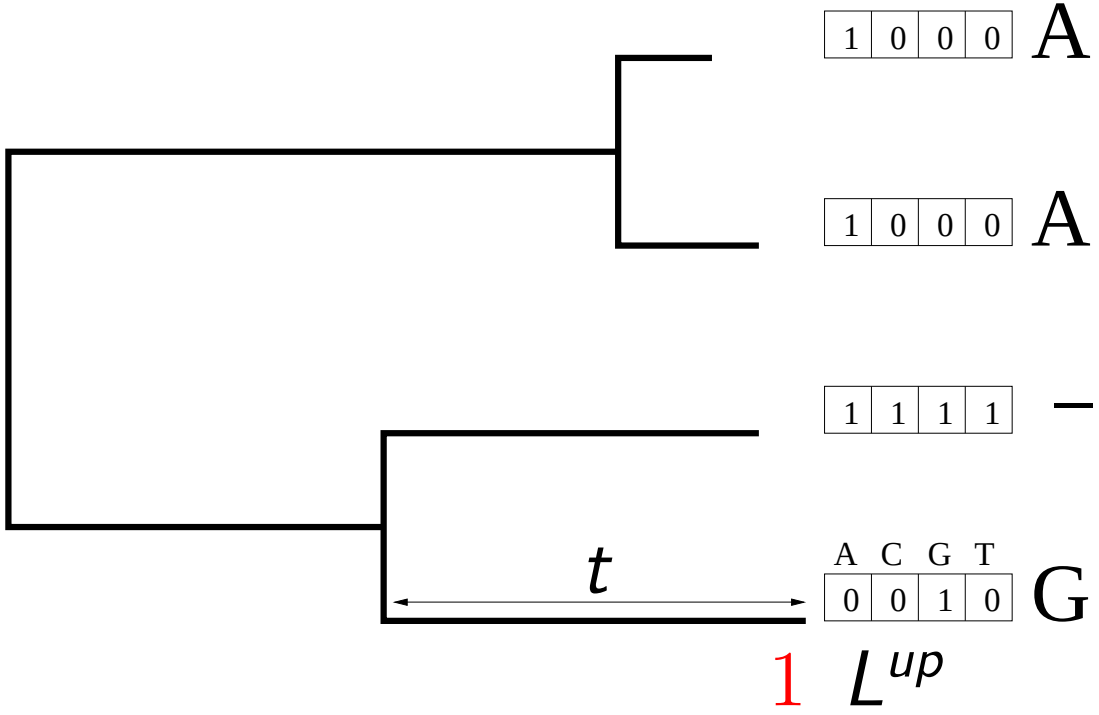


Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)



Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

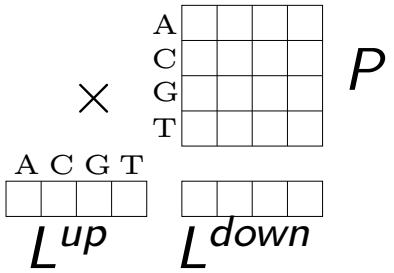
1) $P = e^{tQ}$

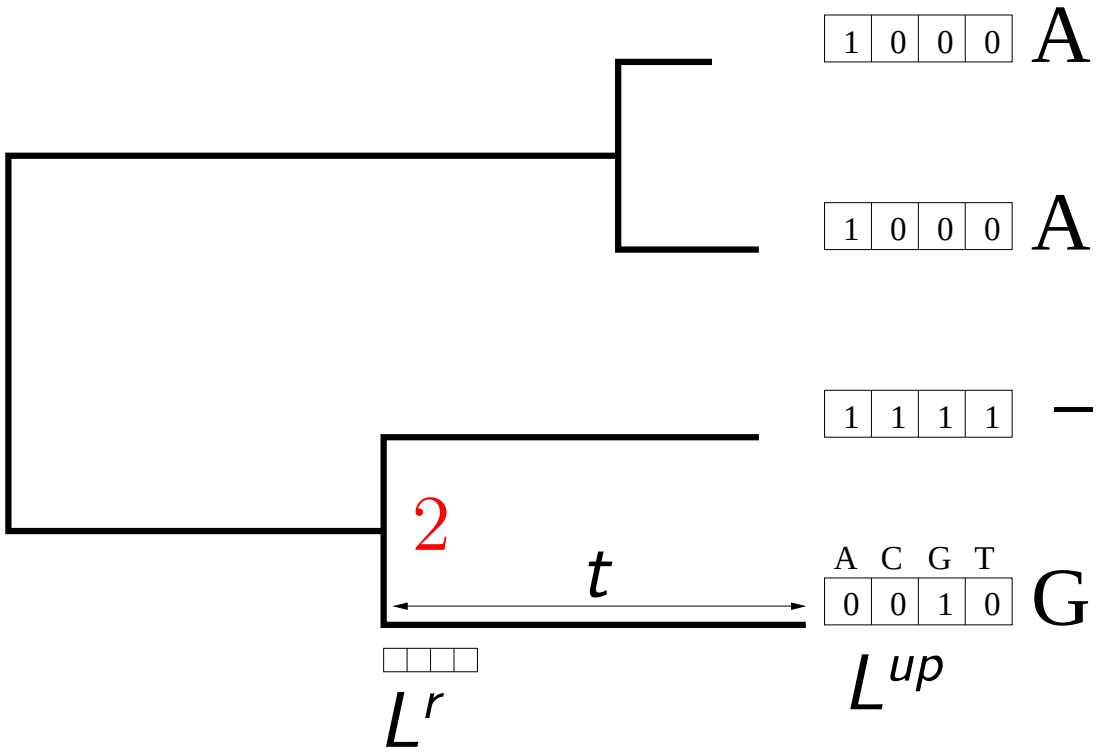


Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

1) $P = e^{tQ}$

2) $L^{down} = L^{up}.P$

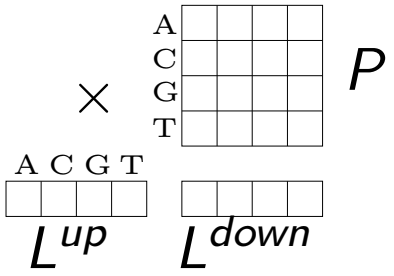


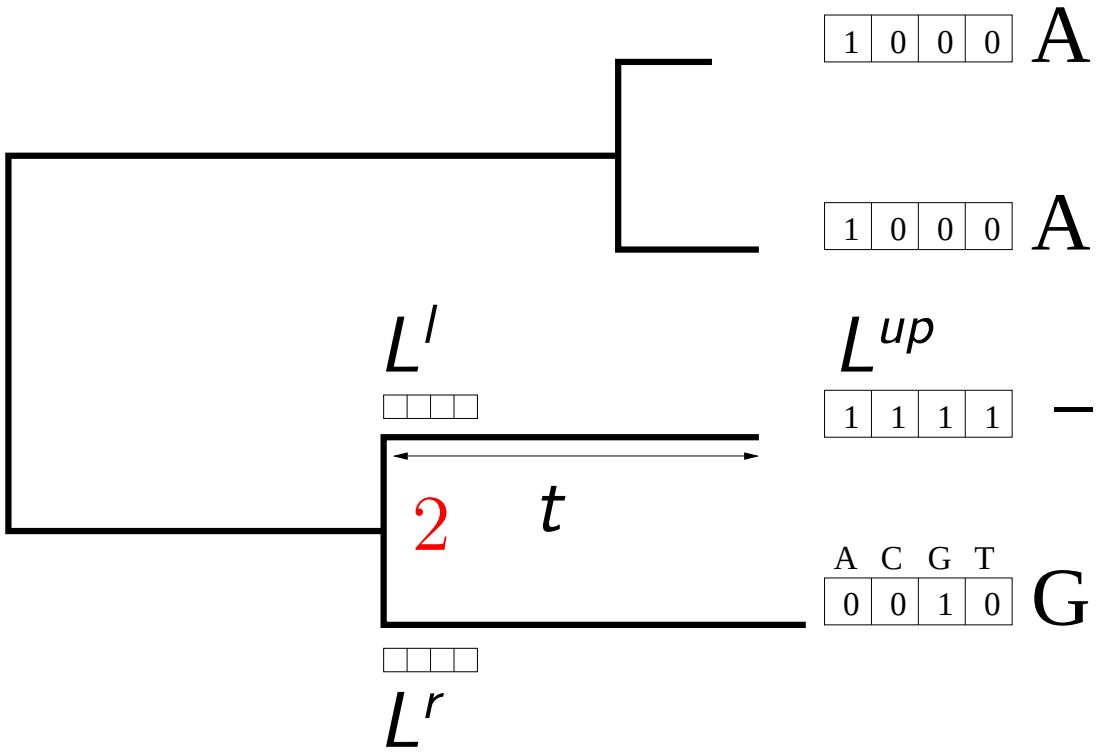


Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

1) $P = e^{tQ}$

2) $L^{down} = L^{up}.P$

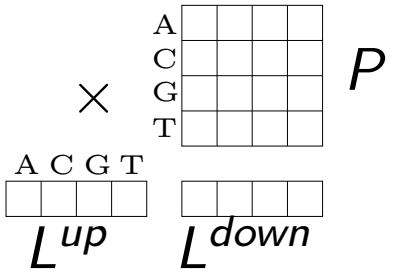




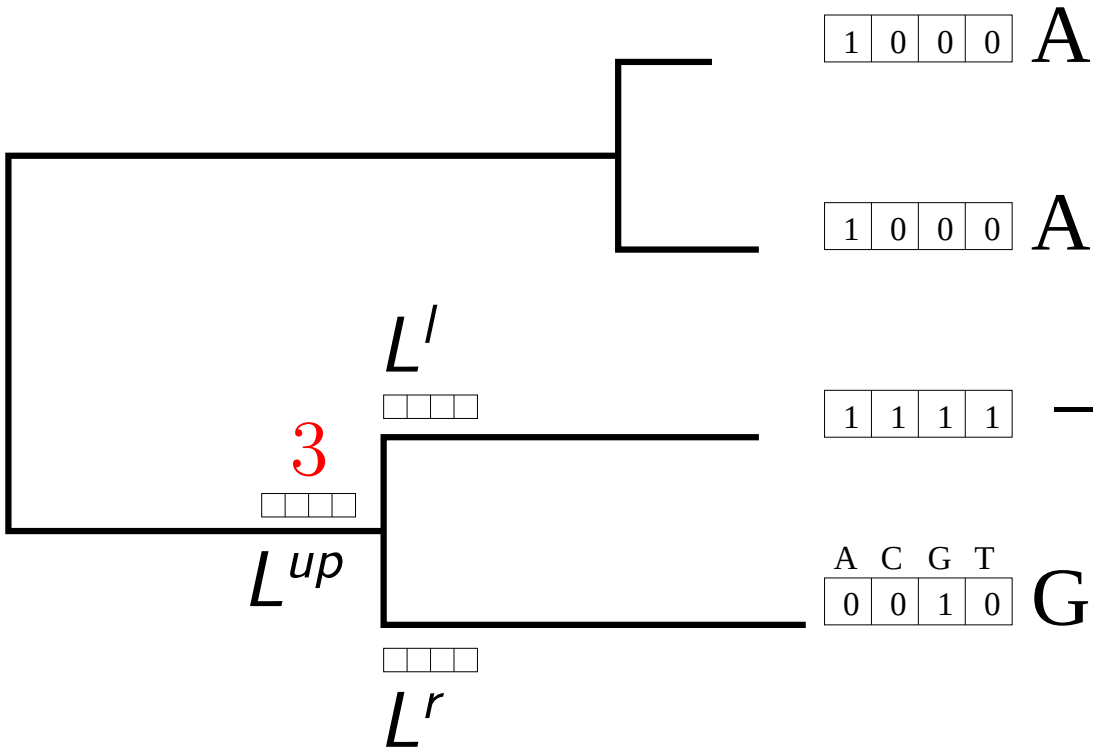
Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

1) $P = e^{tQ}$

2) $L^{down} = L^{up}.P$



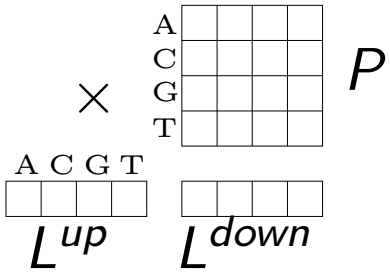
3) $L_i^{up} = L_i^l L_i^r$



Calcul récursif de la vraisemblance, ou probabilité des données sachant les paramètres du modèle (Felsenstein 1981)

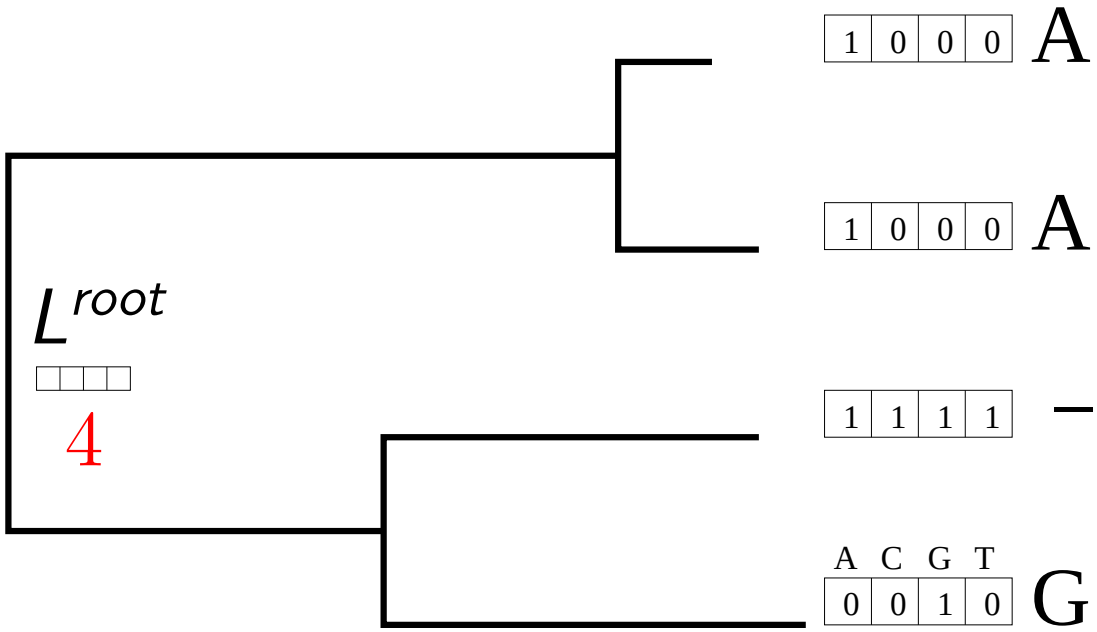
1) $P = e^{tQ}$

2) $L^{down} = L^{up}.P$



3) $L_i^{up} = L_i^l L_i^r$

4) $L = \sum_i \pi_i L_i^{root}$



Partie 3 : Autres caractéristiques de l'évolution moléculaire modélisées en phylogénétique

Sélection, dérive et vitesse d'évolution

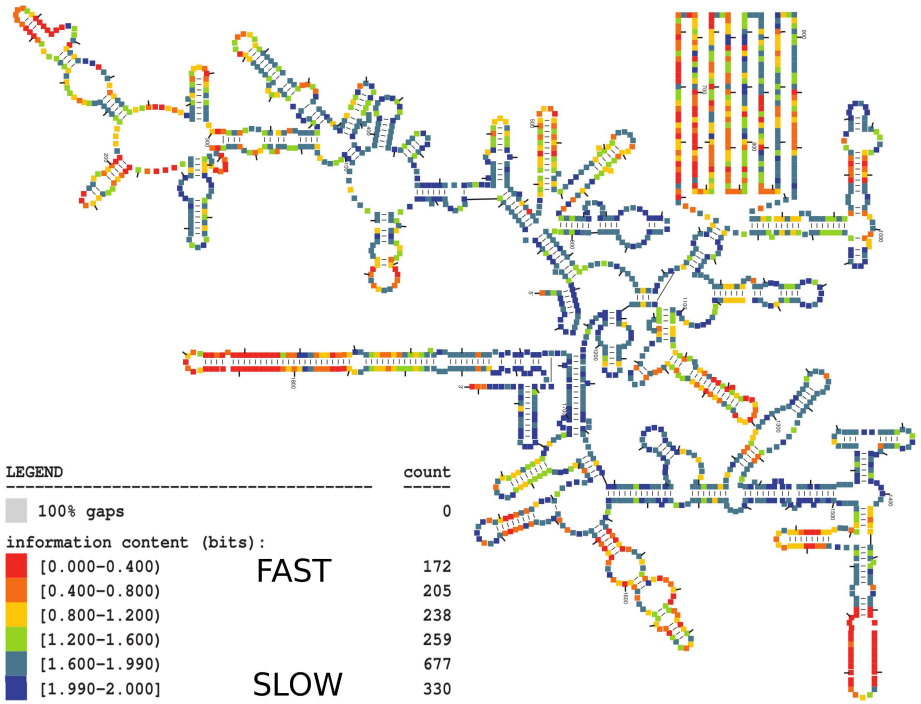
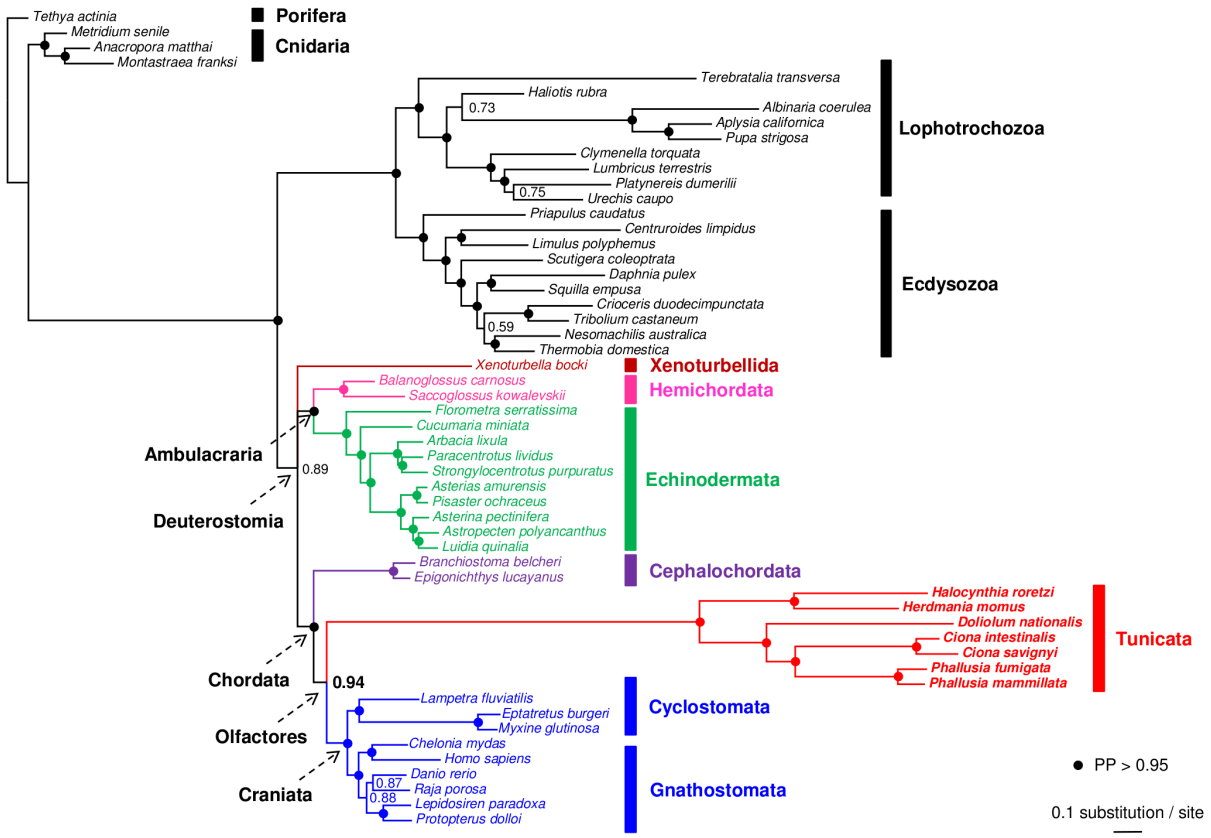


FIGURE: 18S conservation estimated with SSU-ALIGN

- Certains sites évoluent lentement (positions 2 des codons),
- Certains sites évoluent rapidement (positions 3 des codons).

Le modèle RAS ("Rate Across Sites", Yang 1994) permet de considérer N vitesses $r_1 < \ldots < r_N$, déterminées selon une loi gamma discrétisée de paramètre α. Chaque site est préférentiellement décrit par un des processus $r_1 Q, \ldots, r_N Q$.

Sélection, dérive et vitesse d'évolution



- Certaines lignées évoluent lentement (ex. cnidaires),
- Certaines lignées évoluent rapidement (ex. tuniciés),

Chaque longueur de branche représente une vitesse globale d'évolution t_j. On applique un processus $t_j Q$ au long d'une branche, lequel produit t_j substitutions par unité de temps.

Causes des artefacts de reconstruction : saturation

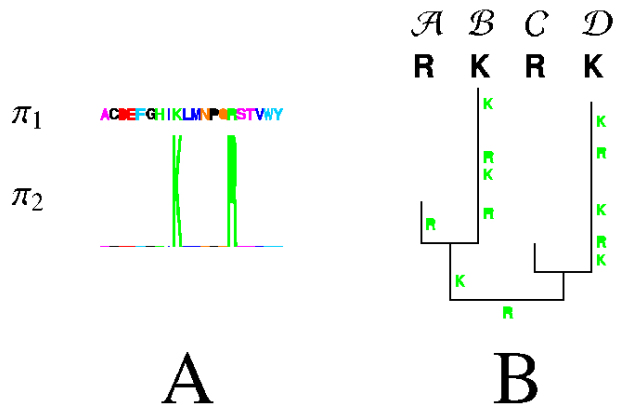


FIG. 7.1: *Illustration théorique d'un site saturé présentant de multiples substitutions exclusivement vers K et R (figure 7.1 B). Ce site appartient de plus à un alignement théorique dont les fréquences empiriques de 20 acides aminés sont égales. Les probabilités stationnaires π_1 d'un processus de substitution partagé par tous les sites sont donc aussi égales entre elles (figure 7.1 A, π_1). Cependant, le site KR considéré est mieux décrit par un processus de substitution dont les probabilités stationnaires π_2 sont importantes exclusivement pour K et R (figure 7.1 A, π_2).*

Le modèle standard sous-estime le risque de convergence pour les sites et/ou les clades rapides. L'artefact couramment observé est l' "attraction des longues branches" (LBA).

- Probabilité avec laquelle le modèle standard estime le risque de convergence au site K/R des espèces $\mathcal{A}$ et $\mathcal{C}$:
  $p \approx \pi_{1,K}^2 = 0.05^2 = 0.0025$,
- Probabilité "réelle" de convergence $p \approx \pi_{2,K}^2 = 0.5^2 = 0.25$.

Causes des artefacts de reconstruction : convergence en composition biochimique

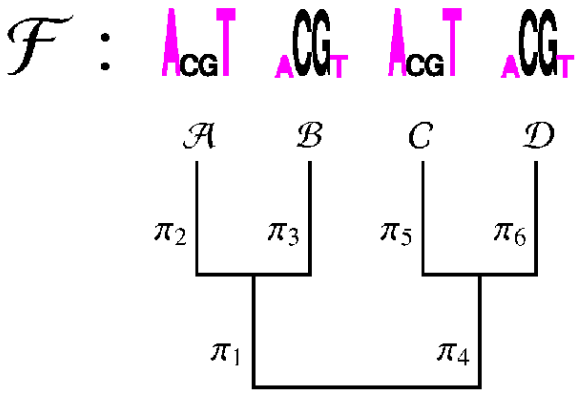


FIG. 6.1: *Illustration théorique d'un jeu de données compositionnellement hétérogène. Les fréquences empiriques, $\mathcal{F}$, de quatre séquences montrent que deux d'entre elles sont enrichies en AT (séquences $\mathcal{A}$ et $\mathcal{C}$), les deux autres en GC (séquences $\mathcal{B}$ et $\mathcal{D}$).*

Le modèle standard sous-estime le risque de convergence pour les sites d'espèce distantes qui dérivent indépendamment vers des compositions similaires (par exemple richesse en AT).

- Probabilité avec laquelle le modèle standard estime le risque de convergence vers A ou T chez les espèces $\mathcal{A}$ et $\mathcal{C}$:
  $p \approx \pi_A^2 = 0.25^2 = 0.0625$,
- Probabilité "réelle" de convergence $p \approx \pi_{2,A} \times \pi_{5,A} = 0.4^2 = 0.16$.

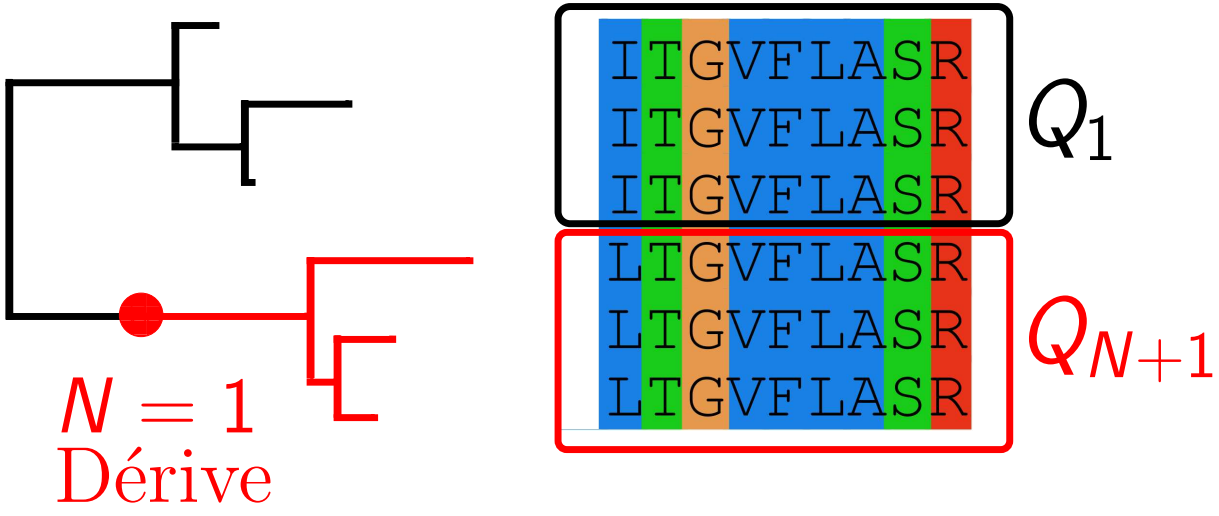
Relaxer l'hypothèse d'homogénéité

Modèle θ

$\theta = \{\tau, \mathbf{t}$

$N,$

$Q_1..Q_{N+1}\}$



- Modélisation de N dérives compositionnelles,
- $N + 1$ processus de substitutions, $Q_1, ..., Q_{N+1}$.

A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. **Blanquart** & Lartillot. Molecular Biology and Evolution (2006).

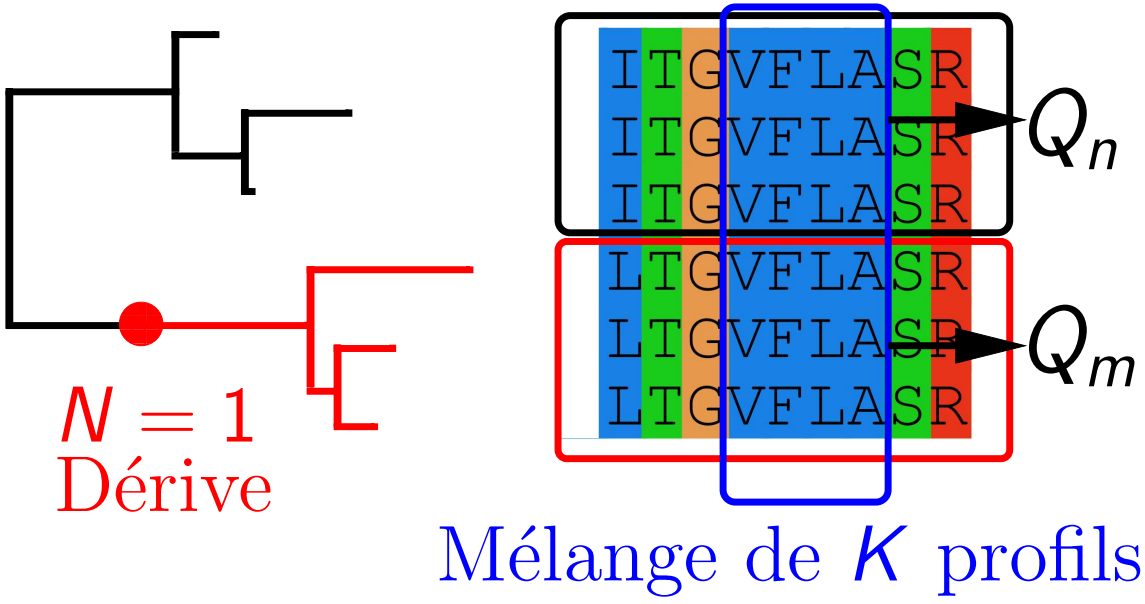
Relaxer l'hypothèse d'homogénéité

Modèle θ

$\theta = \{\tau, \mathbf{t}$

N, K

$Q_1..Q_{K(N+1)}\}$



- Modélisation de N dérives ET de K profils biochimiques,
- $K \times (N+1)$ processus de substitutions, $Q_1, ..., Q_{K\times(N+1)}$.

N **et** K **sont libres et estimés en fonction des données.**

A site- and time-heterogeneous model of amino-acid replacement.

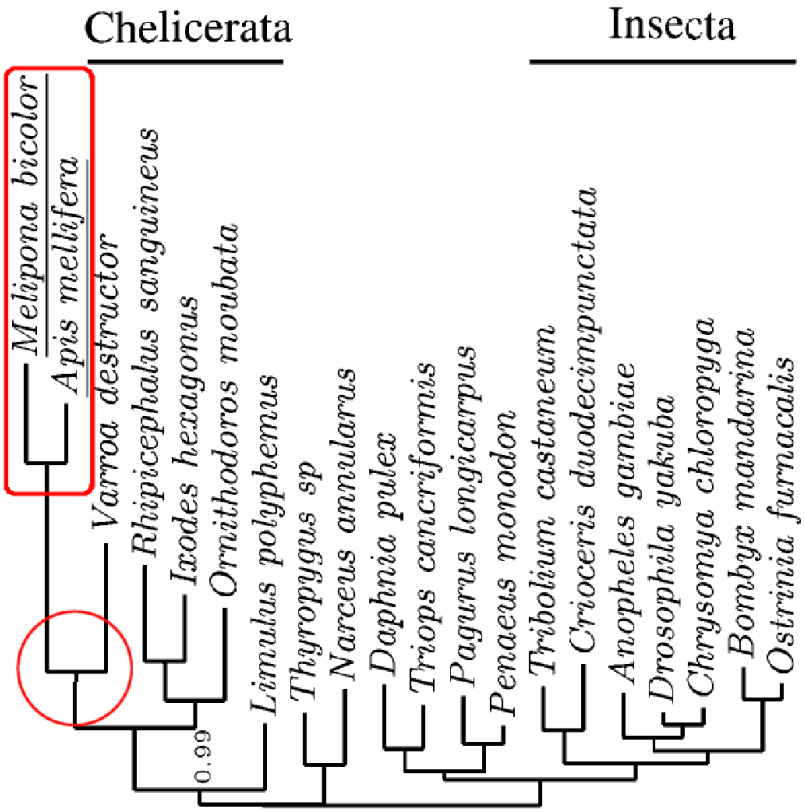
Blanquart & Lartillot. Molecular Biology and Evolution (2008).

Violations des modèles et artefacts de reconstruction

Insectes Apis, Melipona
Tique Varroa

Riches en A+T, et
Evoluant rapidement

Regroupement artéfactuel



Artefact de reconstruction sous le modèle standard : les espèces les plus rapides et les plus riches en AT sont artéfactuellement regroupées.

Causes des artefacts de reconstructions

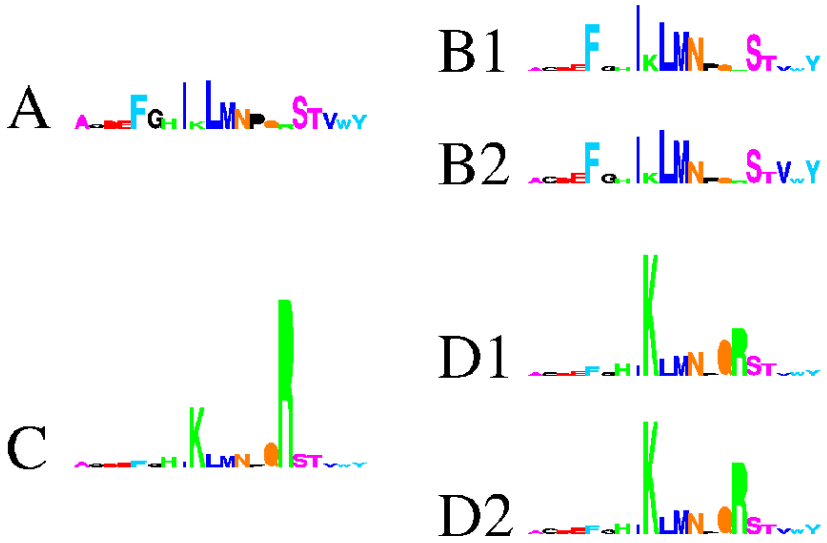


FIG. 7.2: *Probabilités stationnaires moyennes a posteriori inférées par GTR (A), BP (B), CAT (C) et CAT-BP (D), pour l'ensemble des sites exclusivement K et R, et deux branches, l'une menant aux deux espèces d'hyménoptères (B1, D1), l'autre menant à l'espèce Varroa d. (B2, D2). Notons que pour les modèles GTR et BP, les probabilités stationnaires sont les mêmes pour les sites KR et tous les autres sites. De plus, pour GTR et CAT, les probabilités stationnaires sont les mêmes quelles que soient les branches observées.*

Partie 4 : Arbres des espèces et arbres des gènes

Les transferts horizontaux (HGT, Horizontal Gene Transfer)

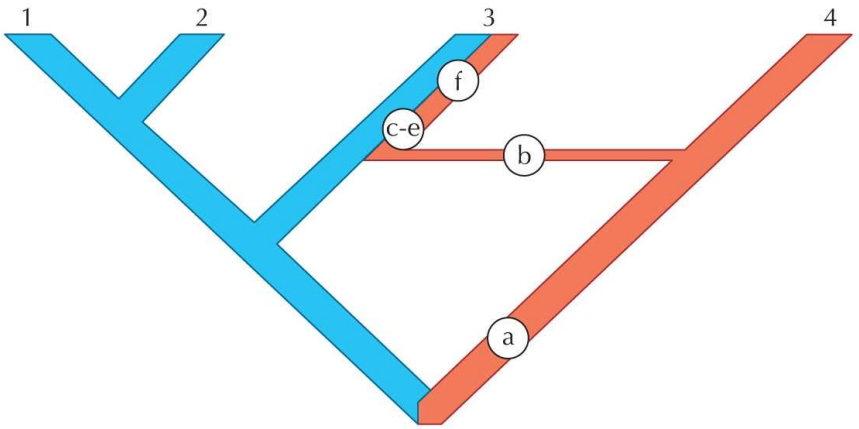
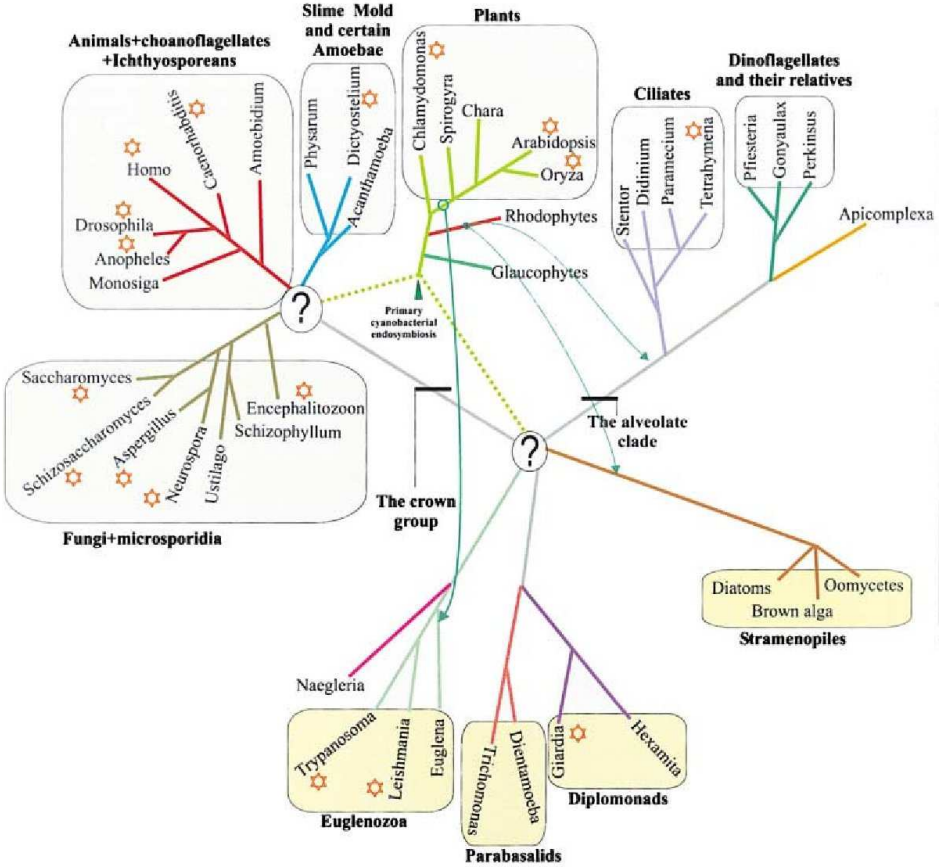


FIGURE 27.30. Stages in lateral gene transfer (LGT). The evolution of four species and one example of LGT are shown. Some key steps in LGT are labeled: (a) Divergence of genomes of different lineages; (b) movement of DNA from one lineage to another; (c) maintenance and replication of the foreign DNA; (d) possible positive selection for the foreign DNA; (e) spread into the new species' population; (f) amelioration. (Modified from Penny D. and Poole A. *Curr. Opin. Genet. Dev.* **9:** 672–677, © 1999 Elsevier.)

Evolution © 2008 Cold Spring Harbor Laboratory Press

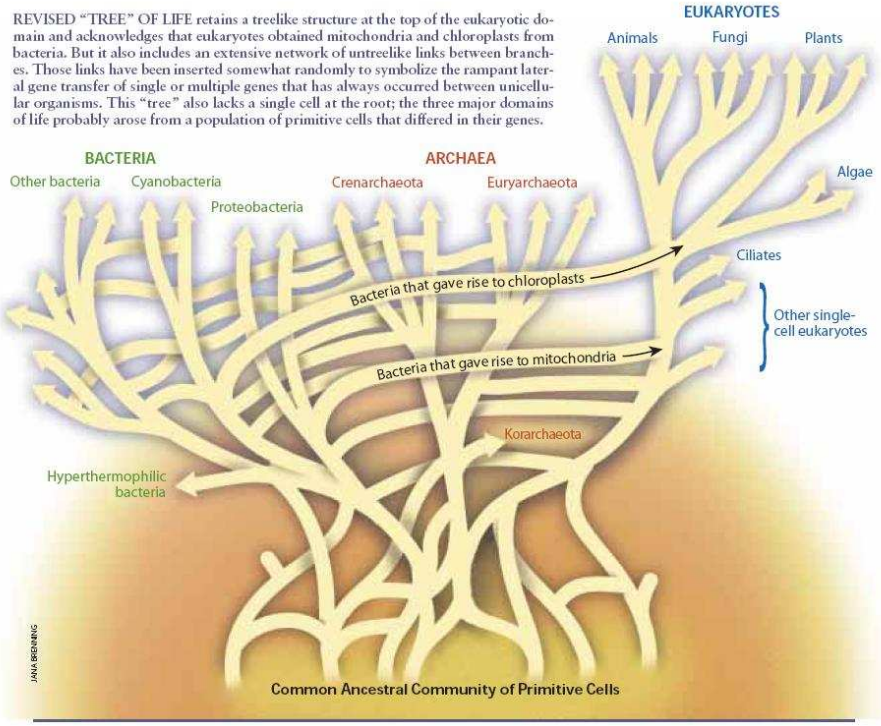
- Le terme horizontal est à opposer à la filiation génétique verticale classique, par reproduction clonale ou sexuée.
- Un transferts horizontal est la captation d'un gène d'une lignée donnée par une autre lignée phylogénétiquement éloignée.

Les transferts horizontaux (HGT, Horizontal Gene Transfer)



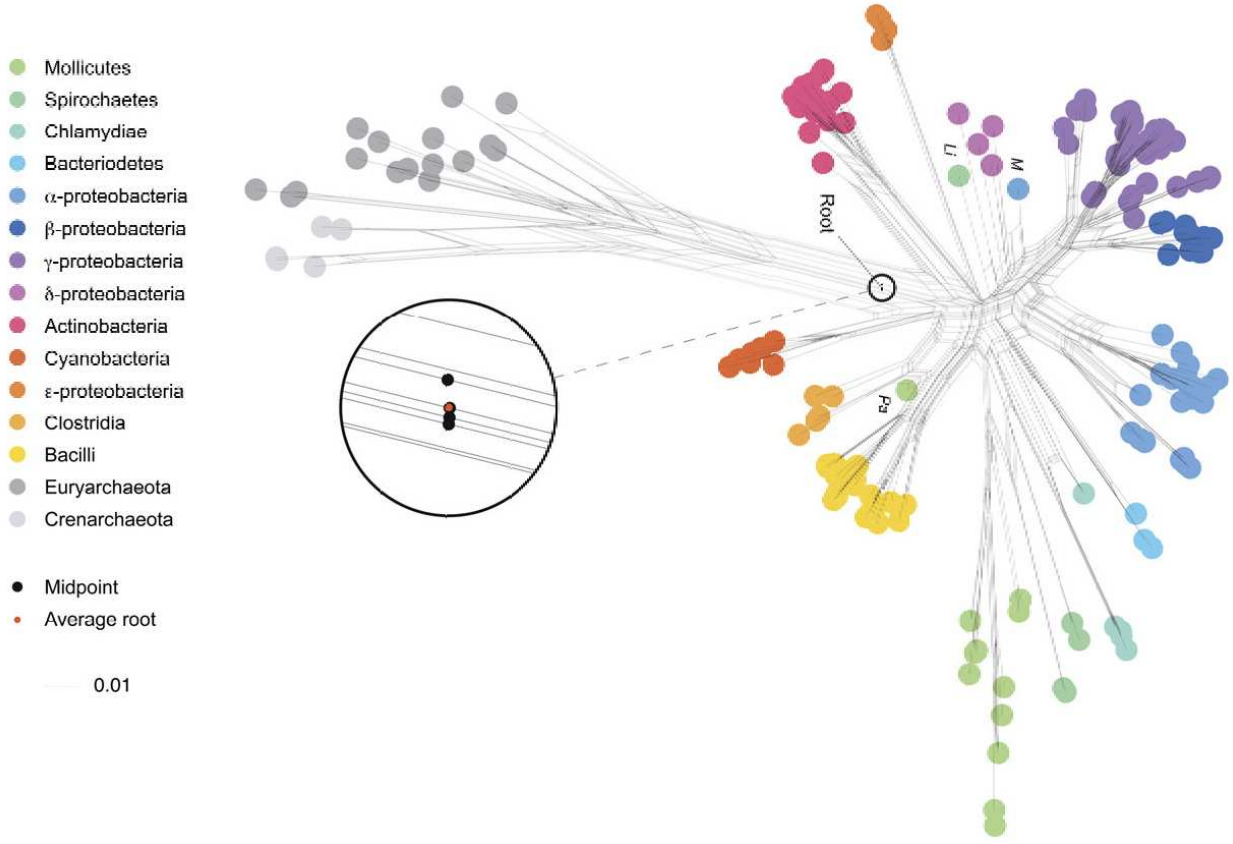
- Les cellules eucaryotes vivent souvent en symbiose et possèdent de nombreux endosymbiontes. Cette proximité éco-physiologique est propice à l'échange horizontal de matériel génétique entre lignées (illustration de Aravind et al 2003, Cell 115, 771-785).

Les transferts horizontaux (HGT, Horizontal Gene Transfer)



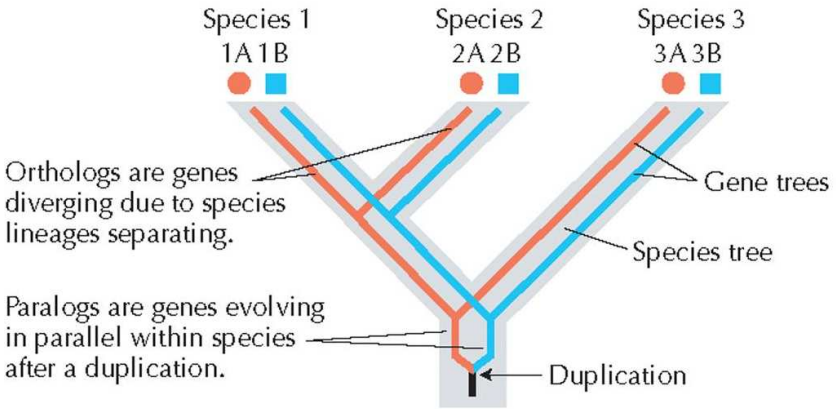
- Eubactéries et Archébactéries échangent également très fréquemment du matériel génétique,
- Certains auteurs doutent de la prééminence de la filiation génétique verticale, et donc du paradigme d'arbre des espèces (Doolittle and Bapteste, PNAS 2006, Pattern pluralism and the Tree of Life hypothesis).

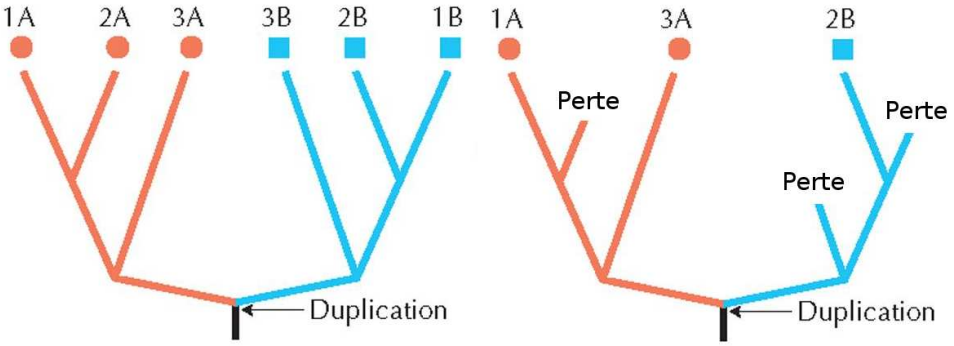
Les transferts horizontaux (HGT, Horizontal Gene Transfer)



- D'autres auteurs montrent que la structure de l'arbre du vivant correspond globalement à celle d'un arbre (Dagan et al. GBE 2010. Genome Networks Root the Tree of Life between Prokaryotic).

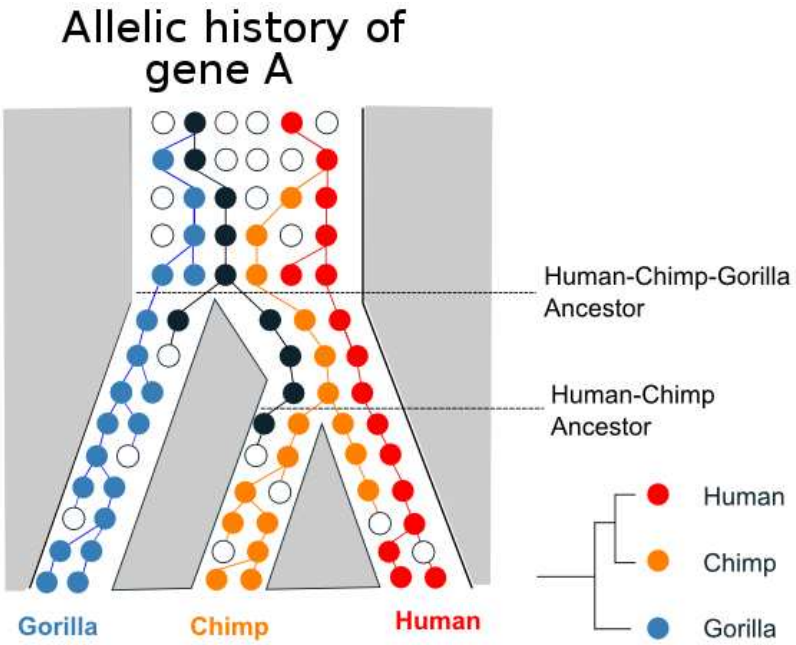
Les duplications et pertes de gènes

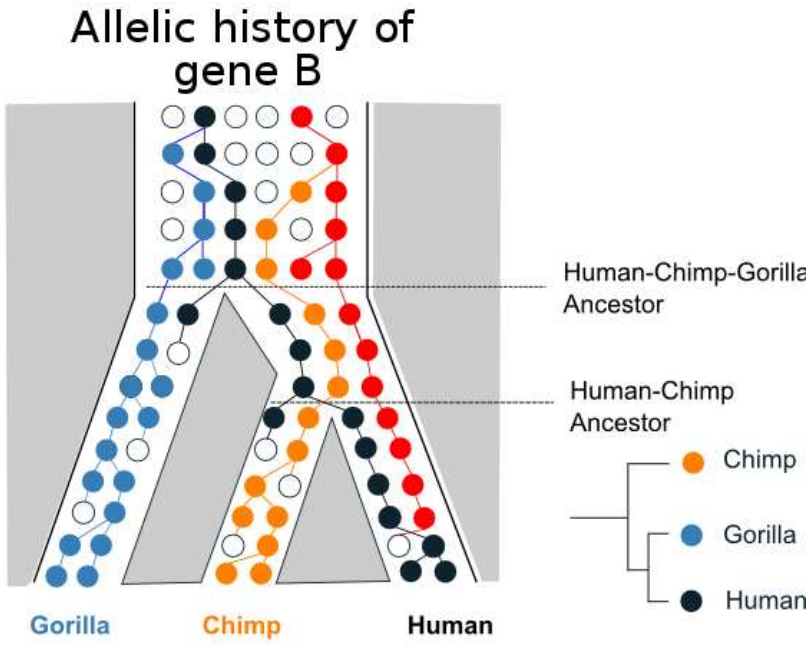




- Chaque spéciation duplique un gène donné dans les deux espèces filles, on parle de deux gènes orthologues
- Des duplications peuvent survenir indépendamment des spéciations, les deux gènes ainsi dupliqués sont dits paralogues.
- Lorsqu'il y a perte de certains paralogues dans certaines lignées, les arbres des gènes et des espèces peuvent apparaître incongruents.
- Les méthodes de "réconciliation" permettent d'estimer un scénario de duplication/perte compatible avec un arbre des espèces.

Génétique des populations, spéciations et tri de lignées incomplets (incomplet lineage sorting)

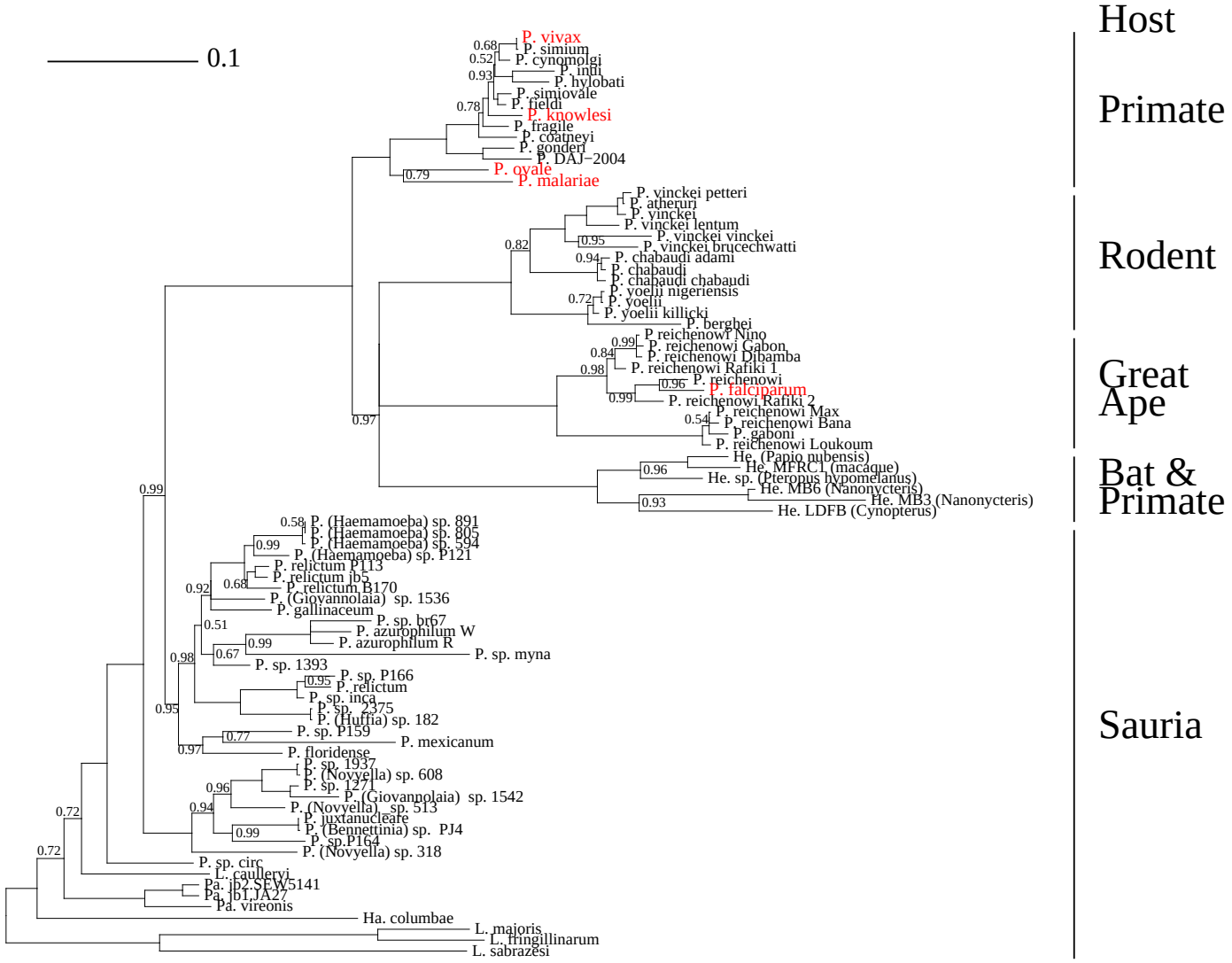




- Pour une même histoire des espèces et lorsque les spéciations sont temporellement proches, les histoires des différentes populations d'allèles peuvent être différentes.
- Ce phénomène, le tri de lignées incomplet, est propre aux espèces à reproduction sexuées.

Partie 5 : Les analyses phylogénétiques comme machines à remonter le temps

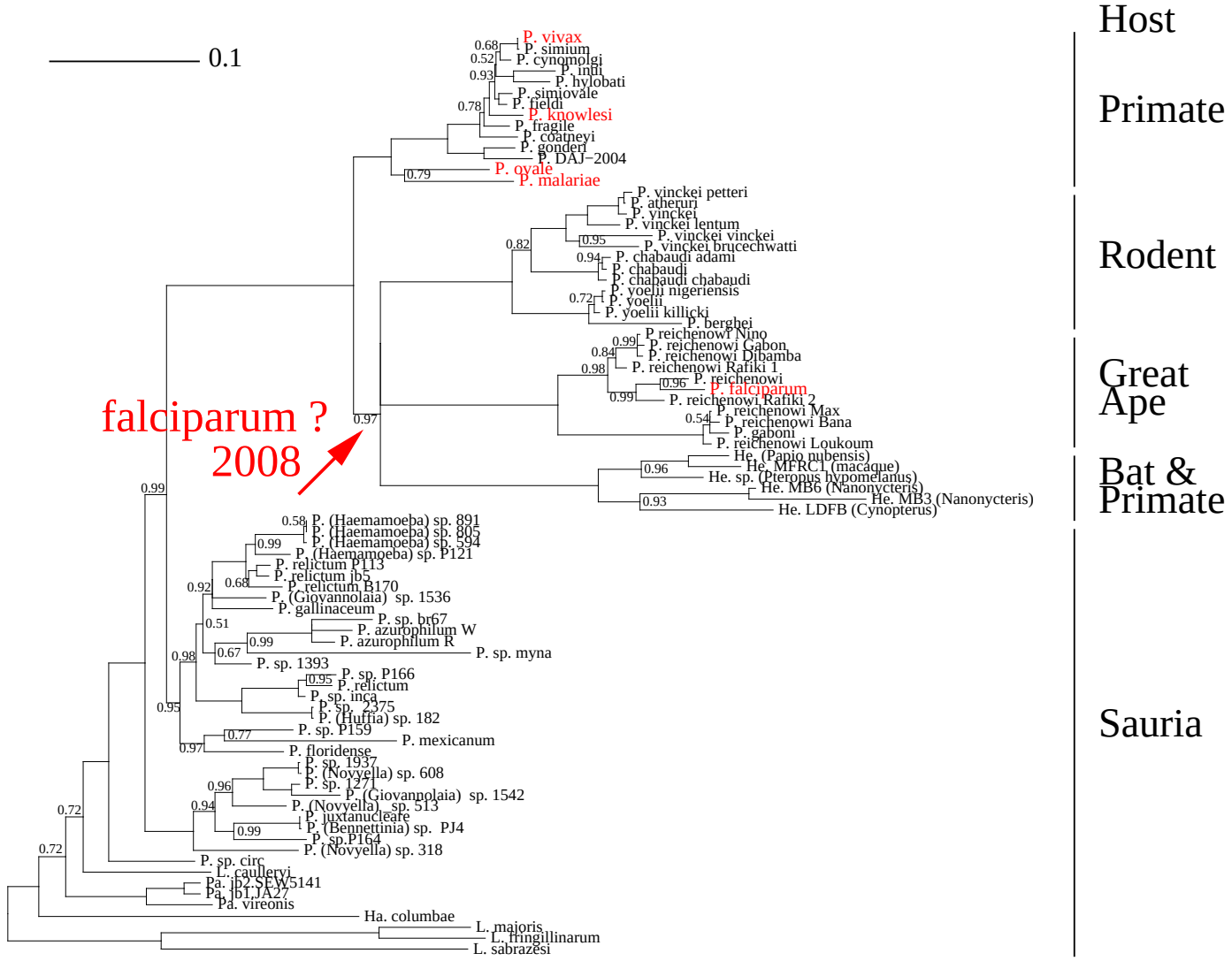
Incertitudes sur la phylogénie des malarias et paléo-écologie



Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes.

Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

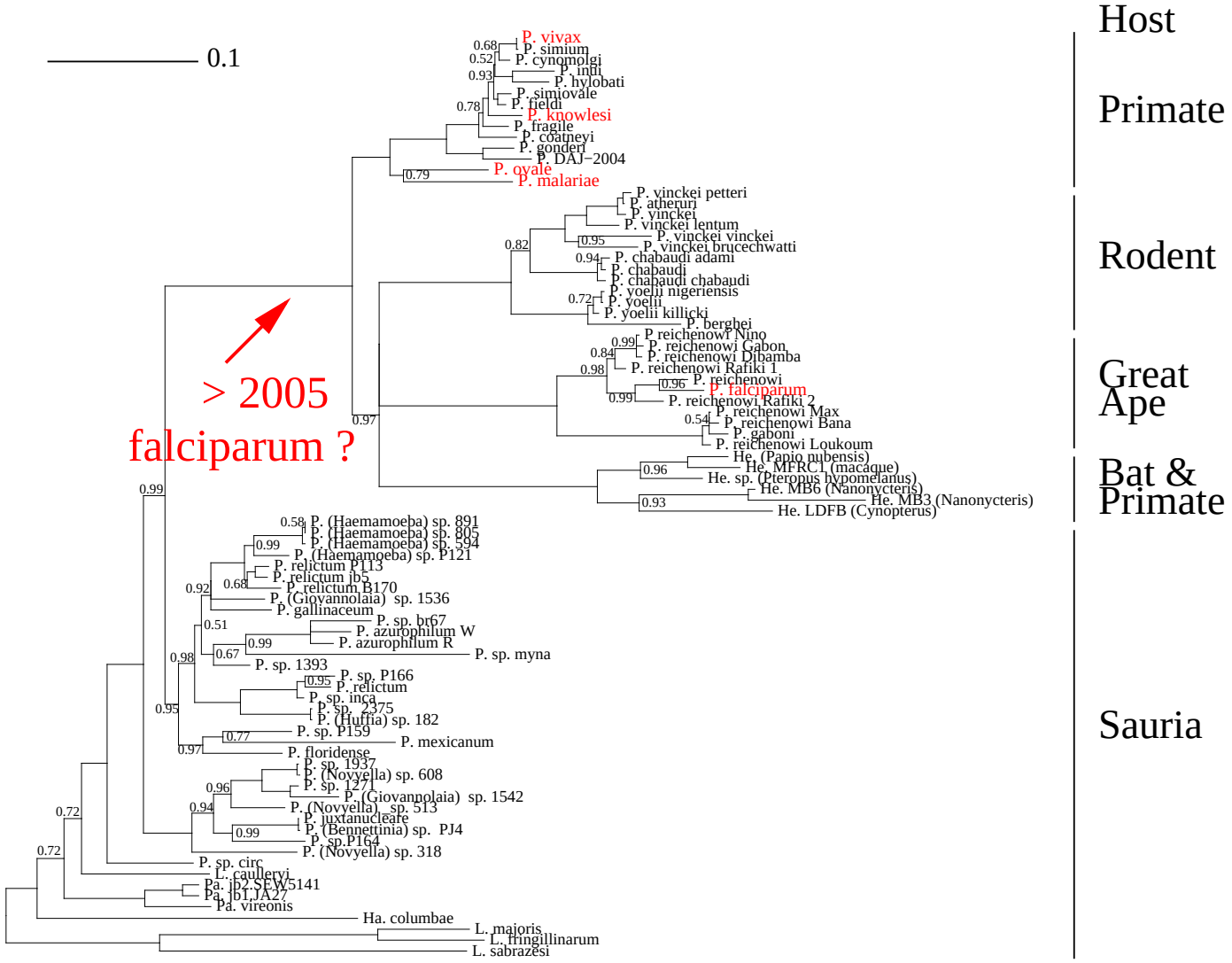
Incertitudes sur la phylogénie des malarias et paléo-écologie



Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes.

Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

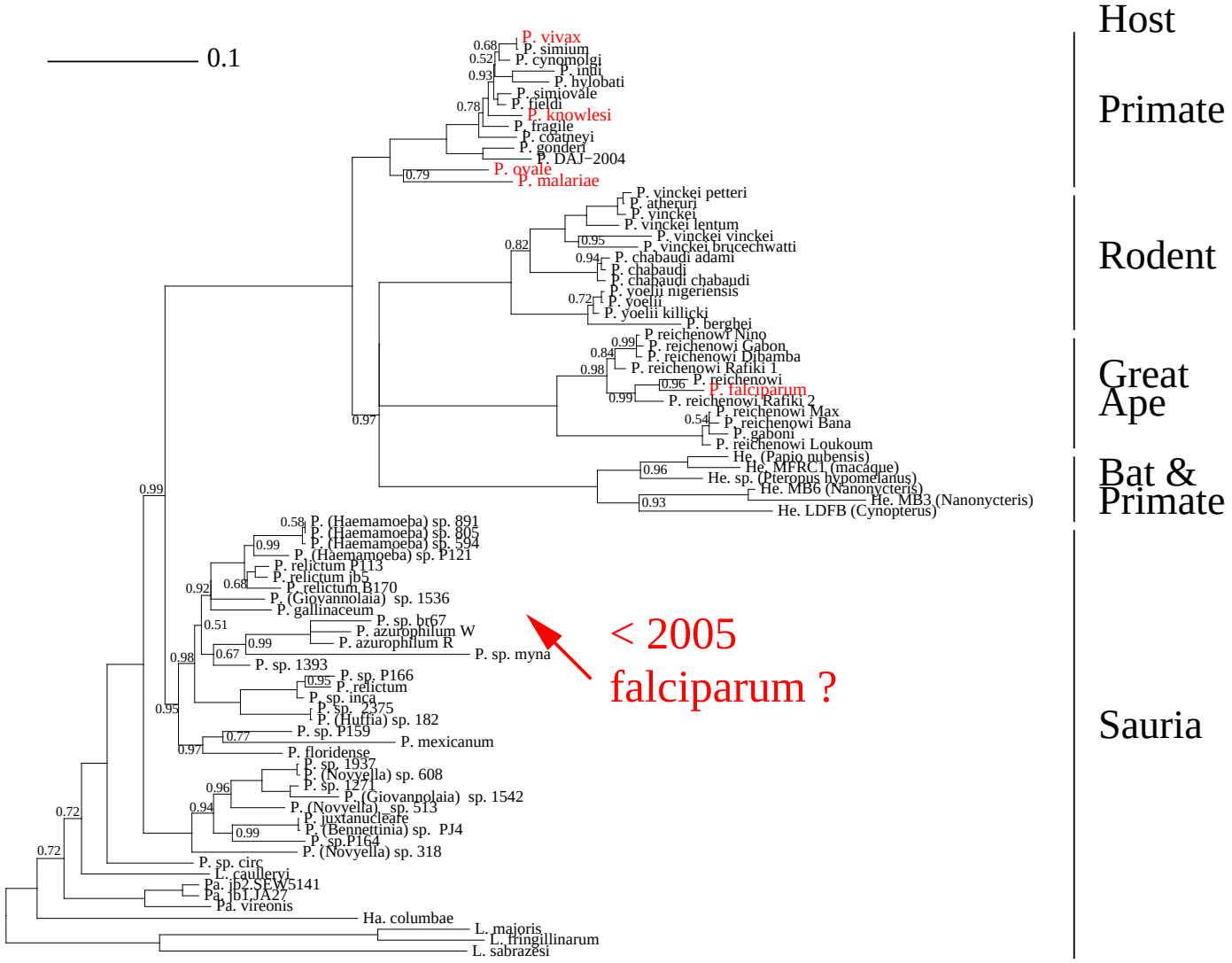
Incertitudes sur la phylogénie des malarias et paléo-écologie



Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes.

Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

Incertitudes sur la phylogénie des malarias et paléo-écologie



Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes.

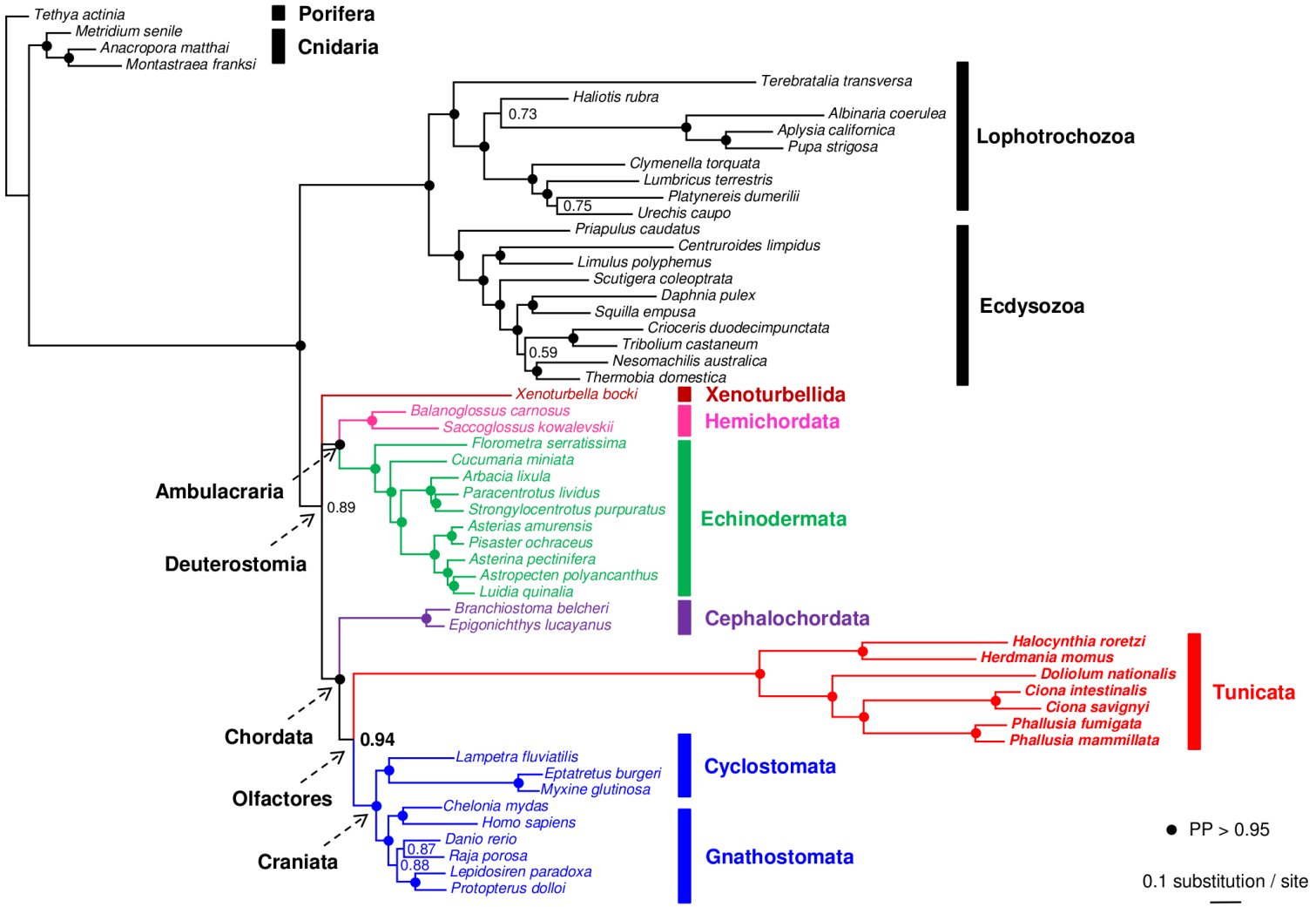
Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

Incertitudes sur la phylogénie des malarias et paléo-écologie

Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes.

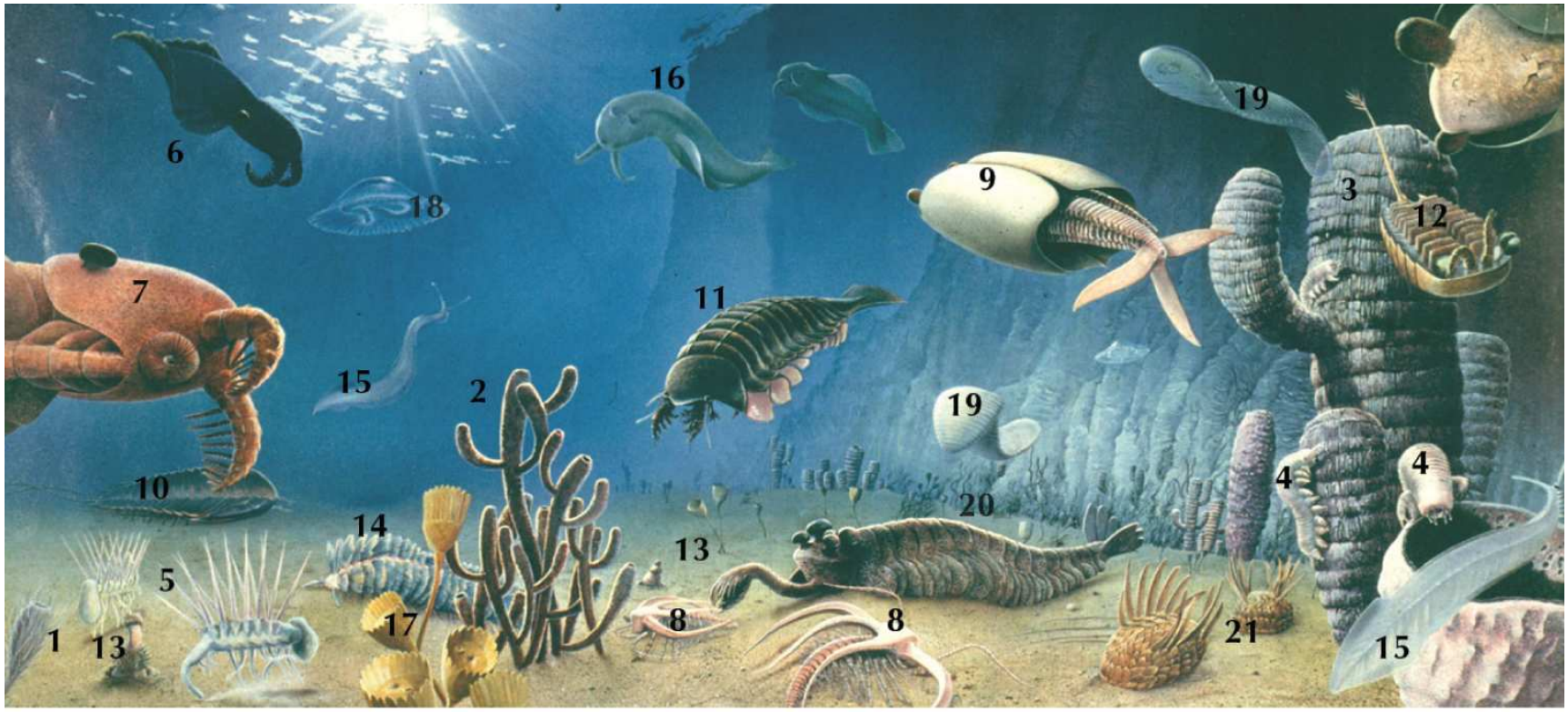
Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

Phylogénie des métazoaires



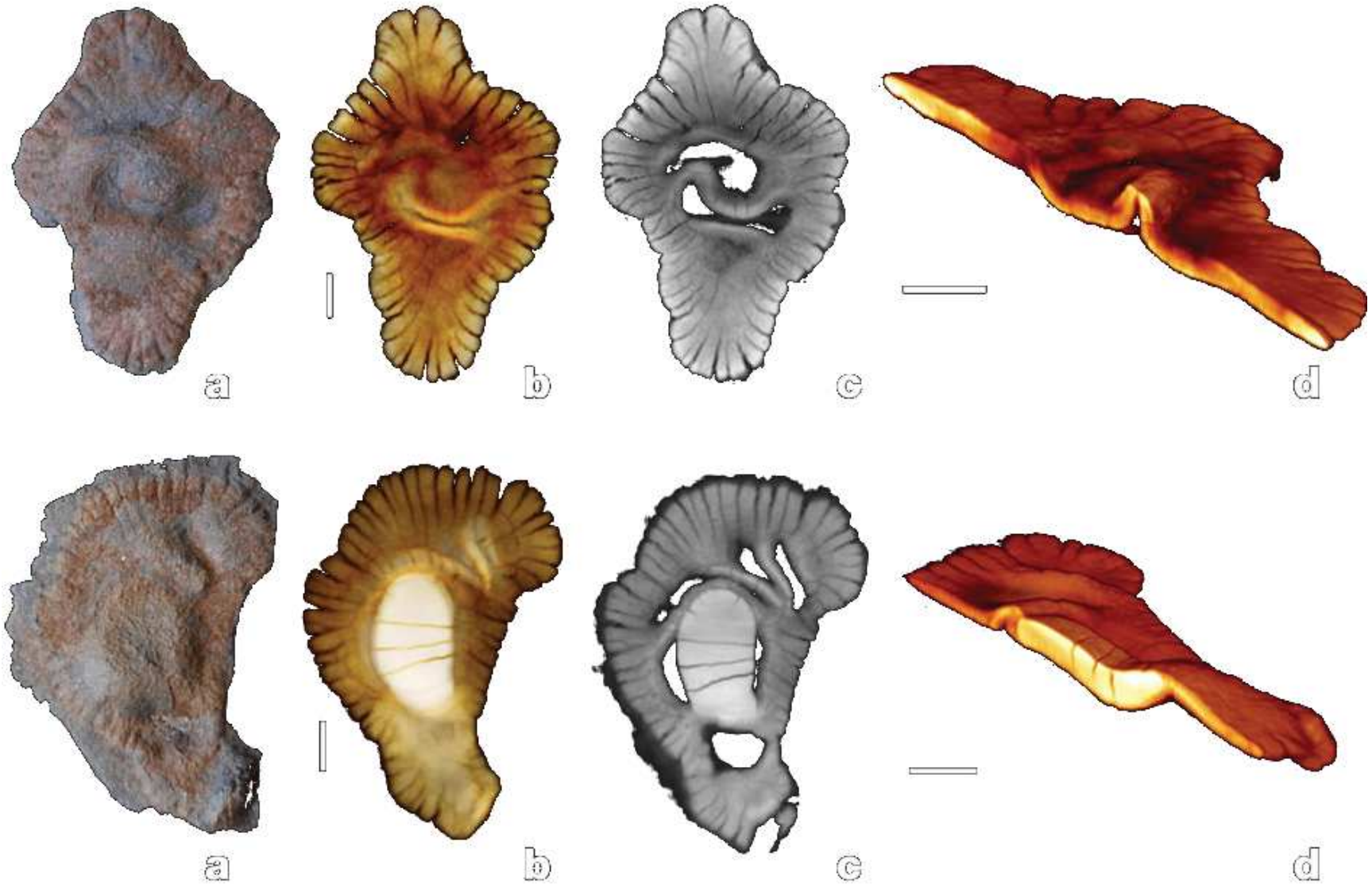
Tunicate mitogenomics and phylogenetics : peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. Singh, Tsagkogeorga, Delsuc, **Blanquart**, Shenkar, Loya, Douzery & Huchon. BMC Evolutionary Biology (2009).

Paléo-faune, Burgess -0.5 (USA) milliards d'années

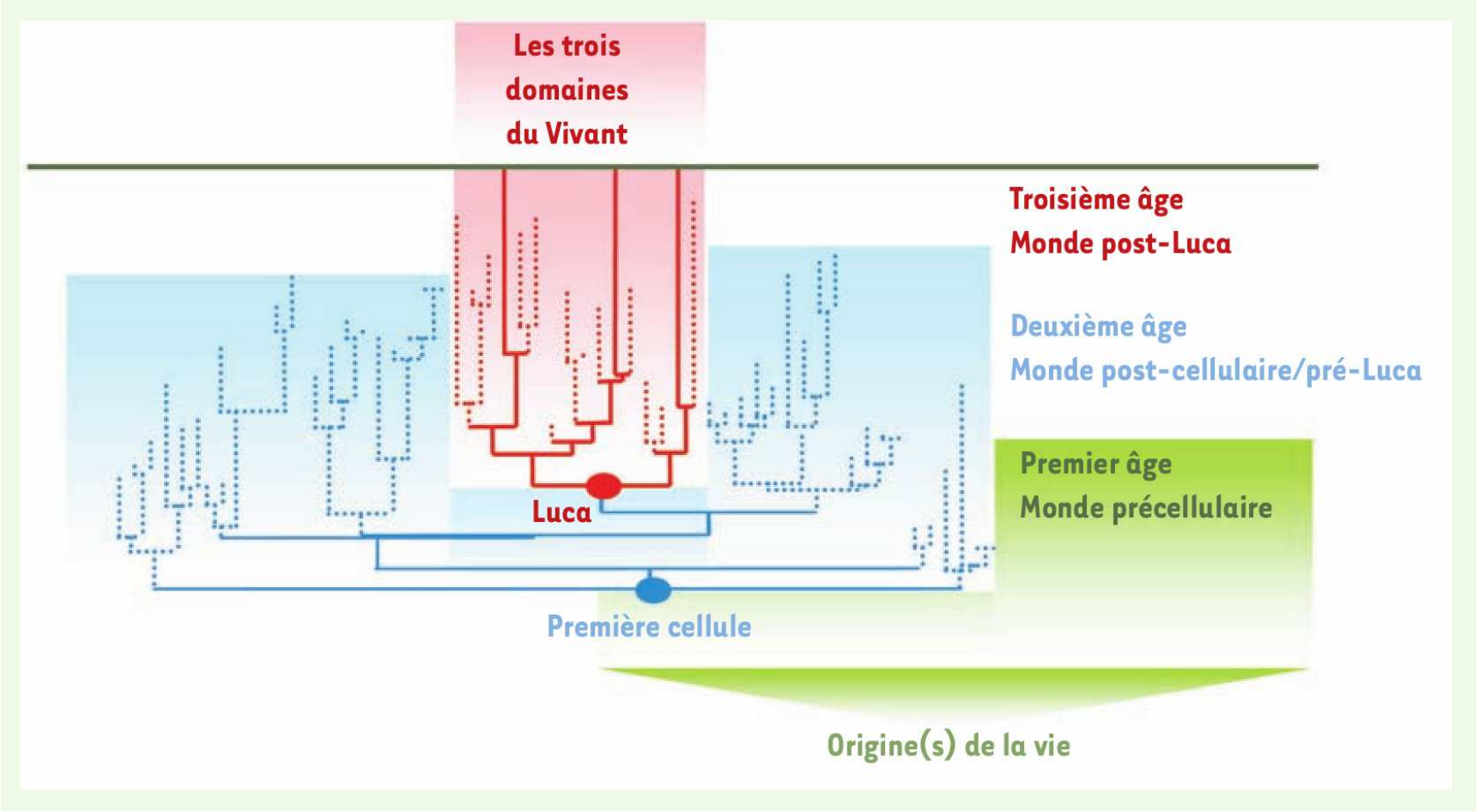


Paléo-faune, Ediacara (Australie) -0.6 milliards d'années

Paléo-faune, Franceville (Gabon), -2.1 milliards d'années



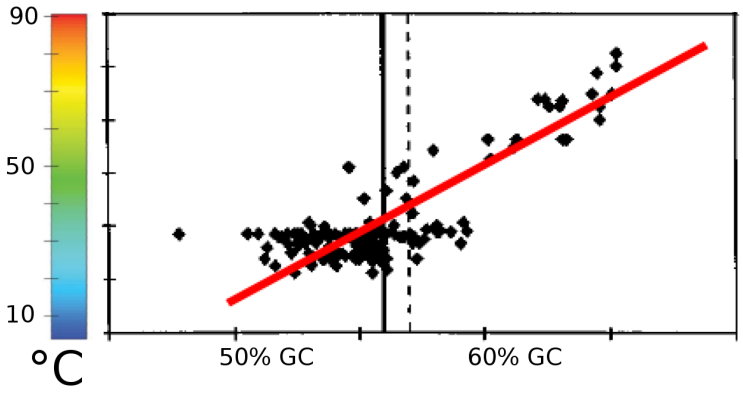
Ecologie des paléo-environnements Archéen et Hadéen



Luca : à la recherche du plus proche ancêtre commun universel. Forterre, Gribaldo & Brochier. Médecine/Science (2005).

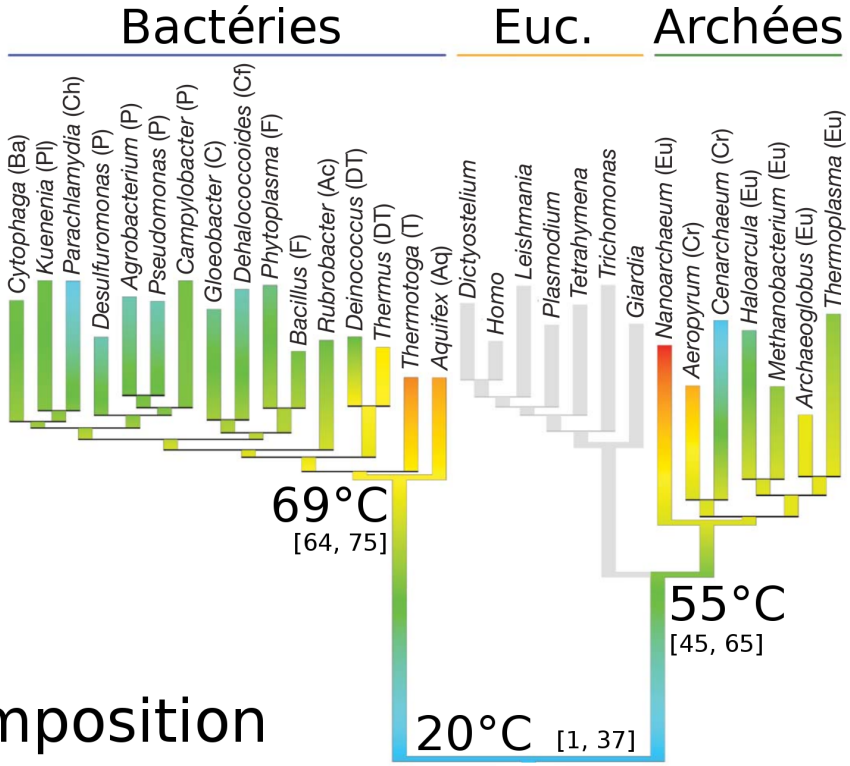
Températures des paléo-environnements

ARN ribosomiques



Galtier, Tourasse, Gouy (1999),
Boussau, Gouy (2006),
Gowri-Shankar, Rattray (2007).

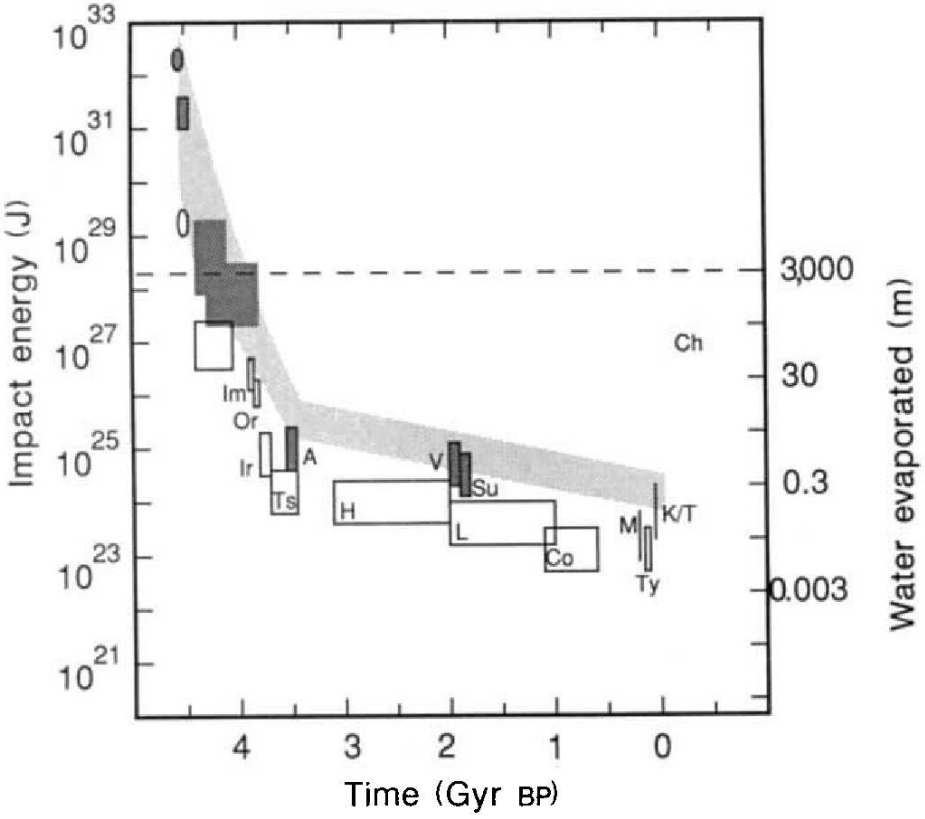
Protéines



Corrélation température/composition
et inférence des compositions ancestrales

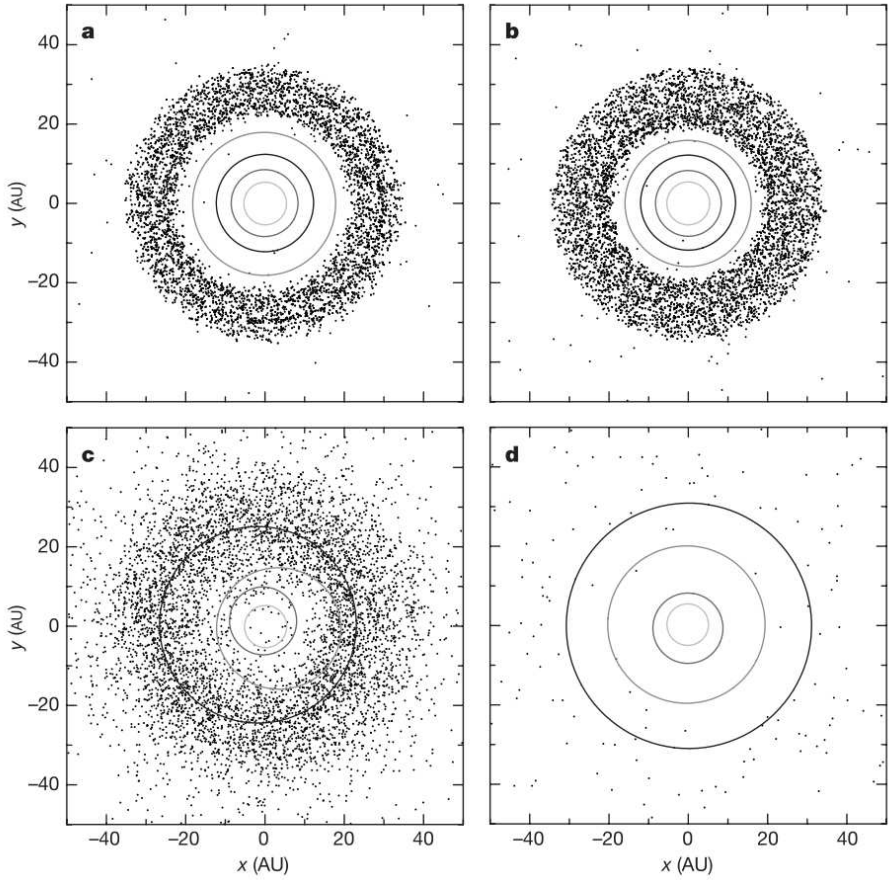
Parallel Adaptations to High Temperatures in the Archean Eon.
Boussau*, **Blanquart***, Necsulea, Lartillot, & Gouy. Nature (2008) (* co premier auteur).

Environnement inter-planétaire



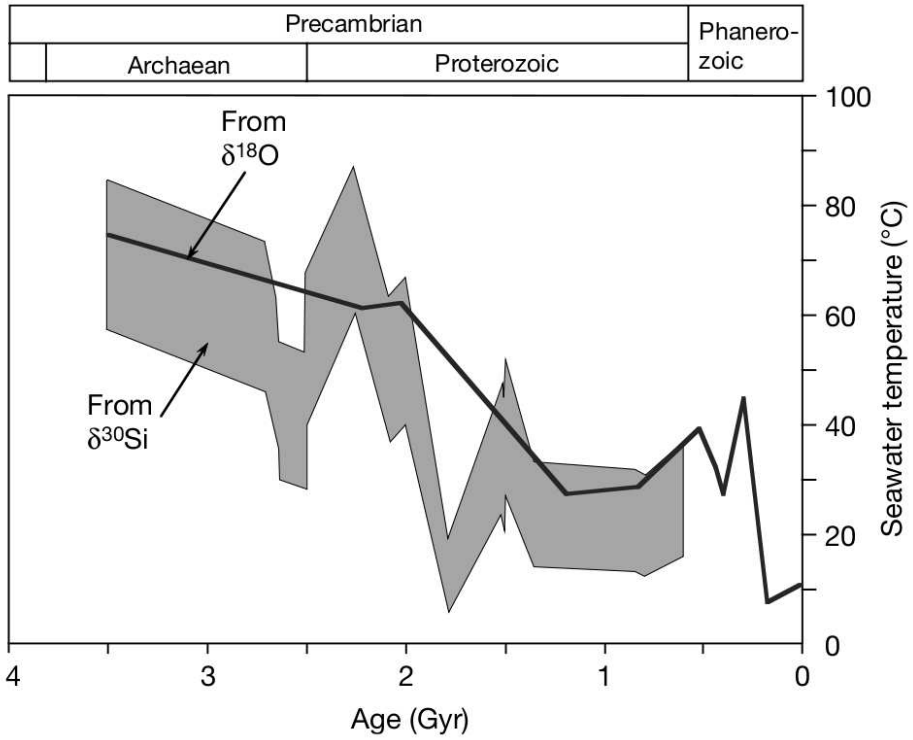
Annihilation of ecosystems by large asteroid impacts on the early Earth.
Sleep, Zahnle, Kasting & Morowitz. Nature (1989).

Le dernier bombardement intense, -3.7 milliards d'années



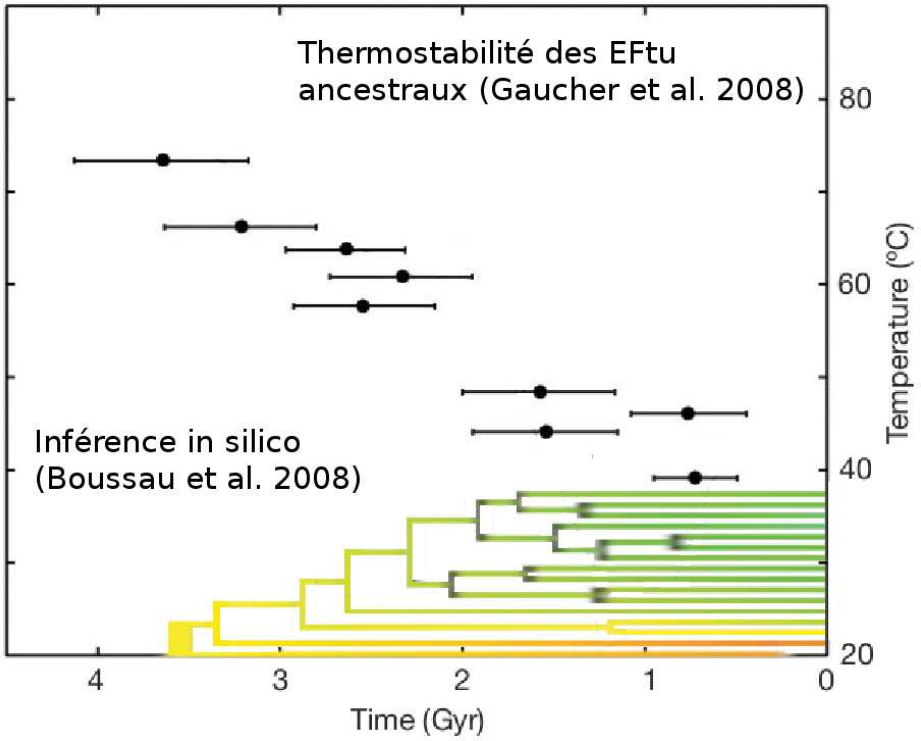
Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Gomes, Levison, Tsiganis & Morbidelli. Nature (2005).

Température des paléo-océans



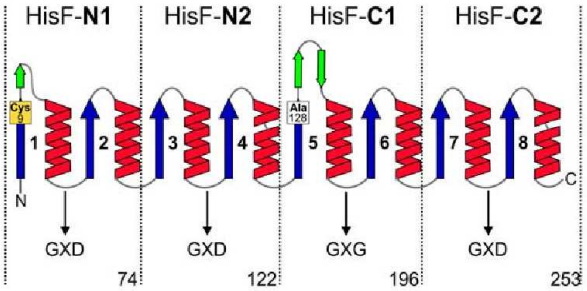
Palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. Robert & Chaussidon. Nature (2006).

Evolution de la thermophilie bactérienne



Palaeotemperature trend for precambrian life inferred from resurrected proteins. Gaucher, Govindara & Ganesh. Nature (2008).

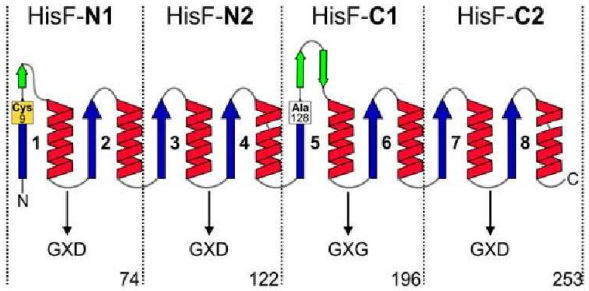
Résurrection de gènes ancestraux



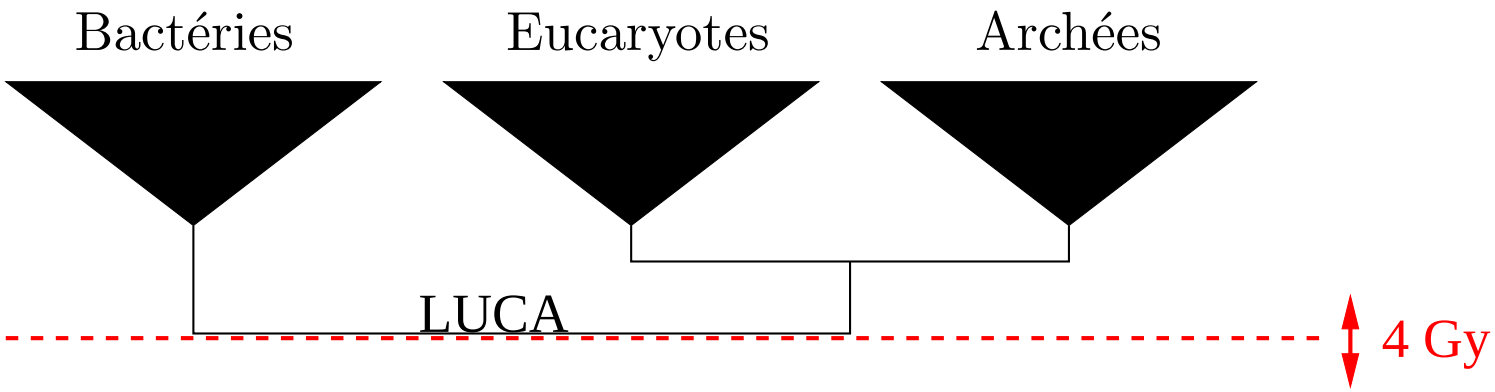
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA

Computational and experimental evidence for the evolution of a $(\beta\alpha)$8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. Richter, Bosnali, Carstensen, Seitz, Durchschlag, **Blanquart**, Merkl & Sterner (2010)

Résurrection de gènes ancestraux



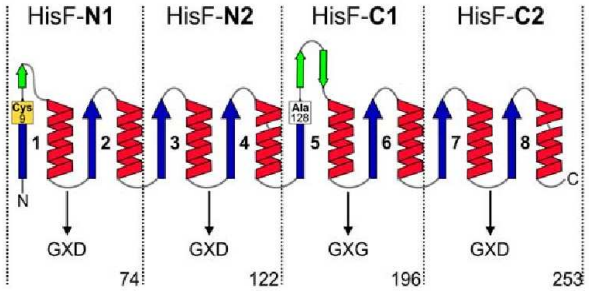
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA



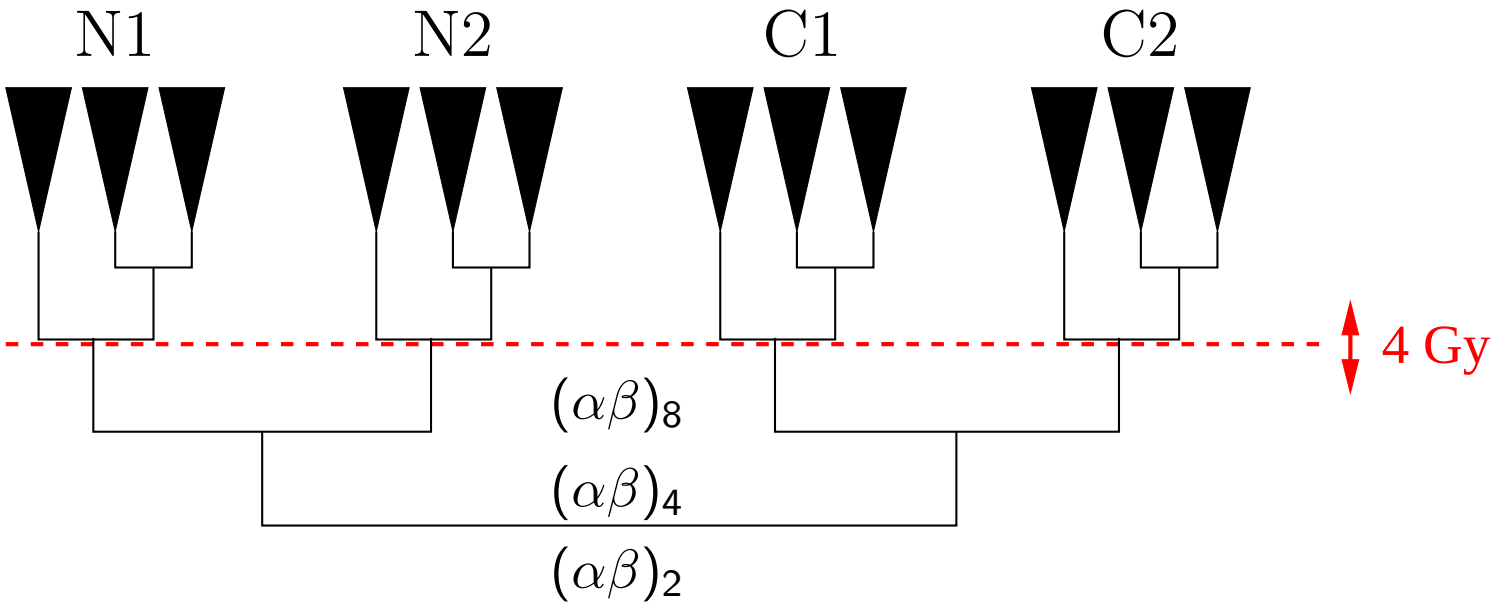
Arbre shématique des 8-mères $(\alpha\beta)_8$

Computational and experimental evidence for the evolution of a $(\beta\alpha)$8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. Richter, Bosnali, Carstensen, Seitz, Durchschlag, **Blanquart**, Merkl & Sterner (2010)

Résurrection de gènes ancestraux



Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA



Arbre shématique des dimères $(\alpha\beta)_2$

Computational and experimental evidence for the evolution of a $(\beta\alpha)$8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. Richter, Bosnali, Carstensen, Seitz, Durchschlag, **Blanquart**, Merkl & Sterner (2010)

Conclusion

- L'information génétique portée par la diversité actuelle du vivant est interprétable par le biais de modèles mathématiques de l'évolution moléculaire, ce qui permet de tirer des conclusions sur des passés parfois remarquablement lointains.
- Le génie génétique permet actuellement de "ressusciter" les séquences moléculaires ancestrales inférées par les modèles, dans le but de les étudier *in vitro* et *in vivo*.
- Les informations délivrées par les analyses phylogénétiques peuvent être corrélées aux données issues de la paléontologie, de la géologie et de l'astrophysique.

- Question : Quelle est la sensibilité des résurrections moléculaires aux incertitudes statistiques liées aux inférences ?
- Question : Quelle est la fiabilité statistique des inférences extrêmement profondes ?
- Question : Quel est "l'horizon entropique" au delà duquel tout signal phylogénétique apparaît trop dégradé pour être interprétable ?