Les modèles phylogénétiques comme machines à remonter le temps

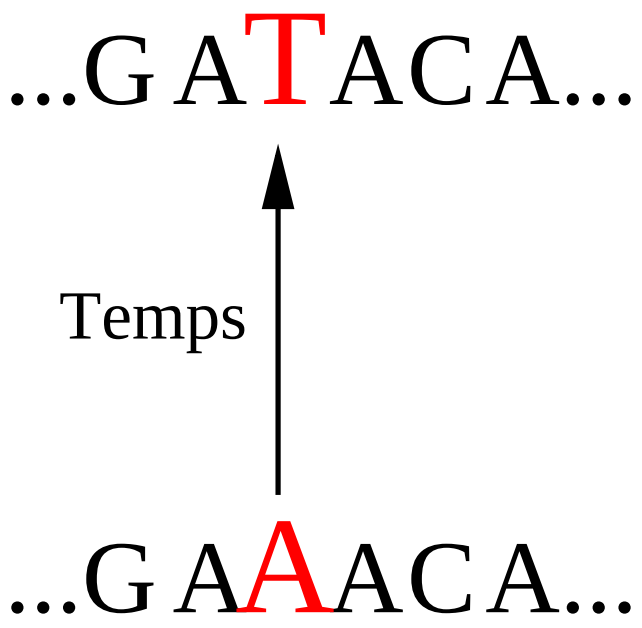
Samuel Blanquart, CR2 INRIA, équipe/projet SEQUOIA2

November 29, 2010

Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution.
Theodosius Dobzhansky.

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:

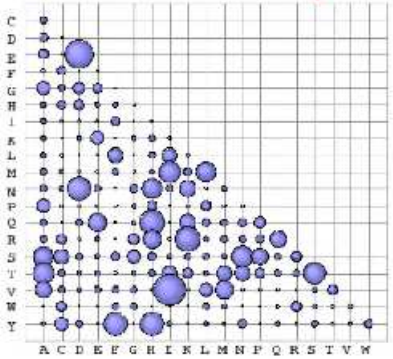
$$P(i \rightarrow j | t) = [e^{t \times Q}]_{i,j}$$

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

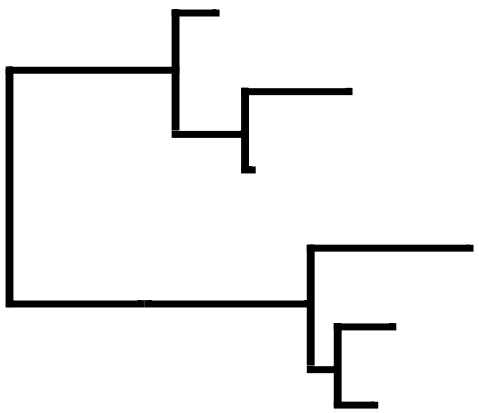


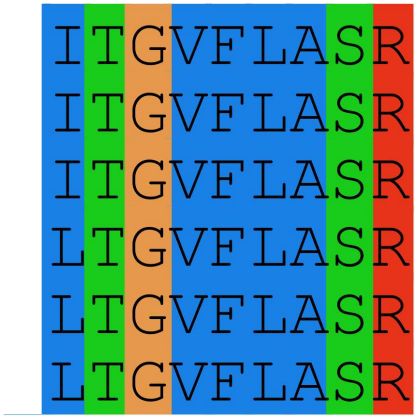
$$\begin{cases} Q_{l \neq m} = \pi_m \rho_{lm} \\ Q_{ll} = -\sum_{m \neq l} Q_{lm} \end{cases}$$

Le modèle standard

Modèle θ

$\theta = \{\tau,$





Données D

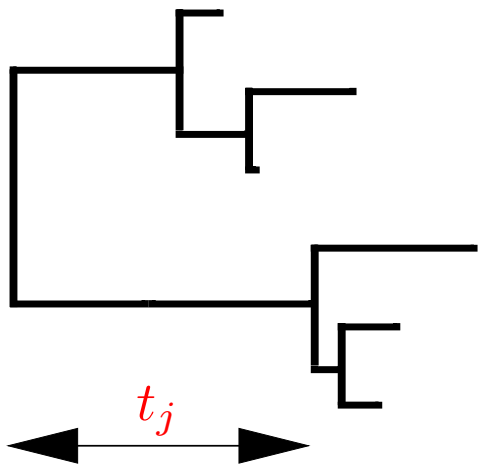
- Données D, ensemble de séquences homologues,
- Modèle: Topologie τ

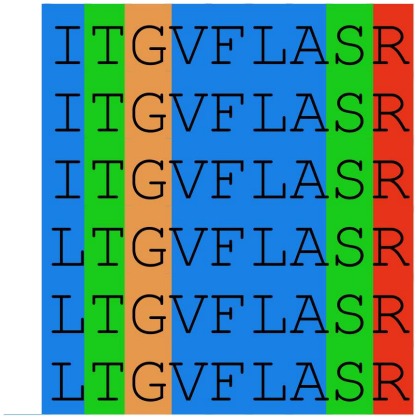
Le modèle standard

Modèle θ

$\theta = \{\tau,$

$\mathbf{t},$





Données D

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

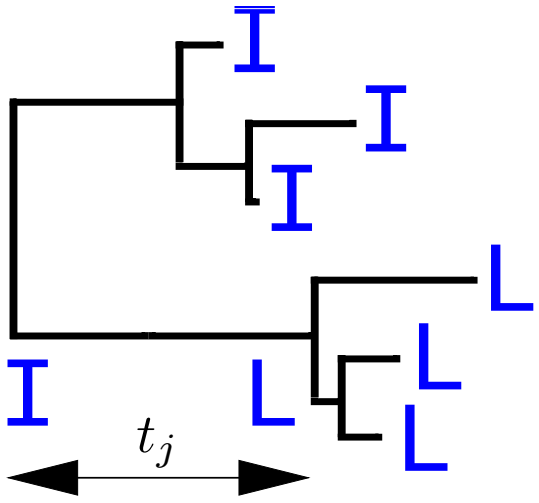
Le modèle standard

Modèle θ

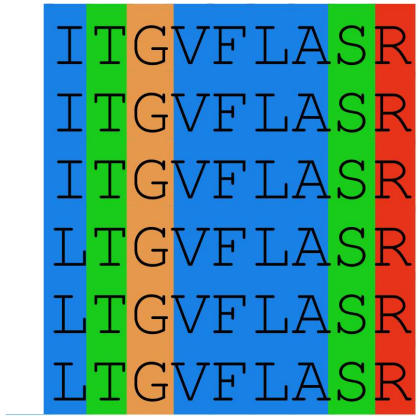
$\theta = \{\tau,$

$\mathbf{t},$

$Q\ \}$



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



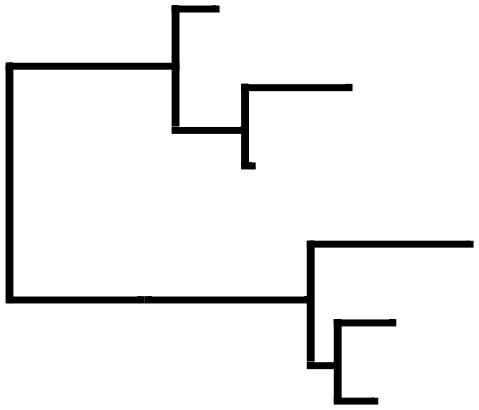
Données D

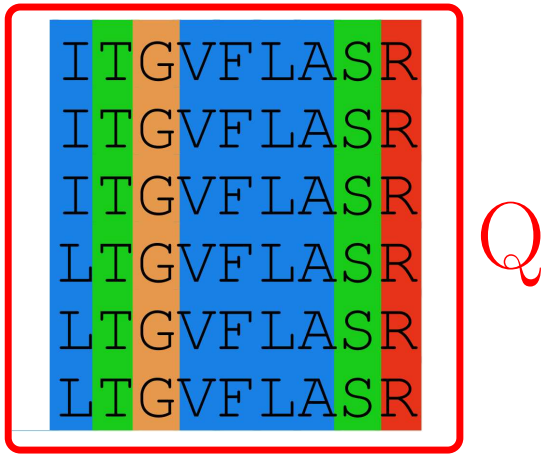
- Données D, ensemble de séquences homologues,
- Modèle: Topologie τ, Vitesses d'évolution $\mathbf{t}$,
- Q, générateur Markovien du processus de substitution.

Le modèle 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.

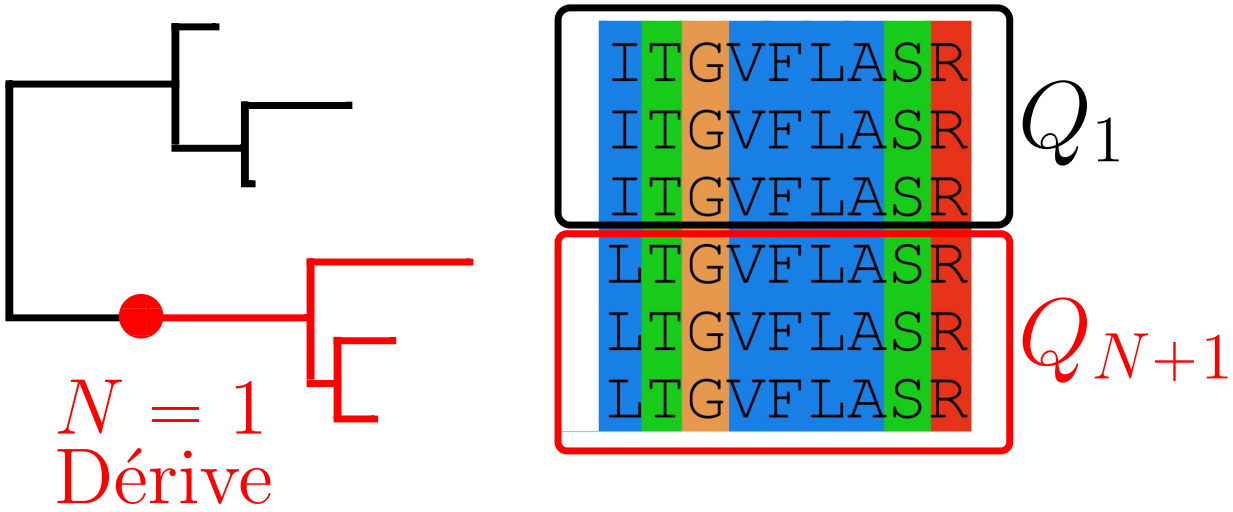
Contribution 1: Relaxer l'hypothèse d'homogénéité

Modèle θ

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

$N,$

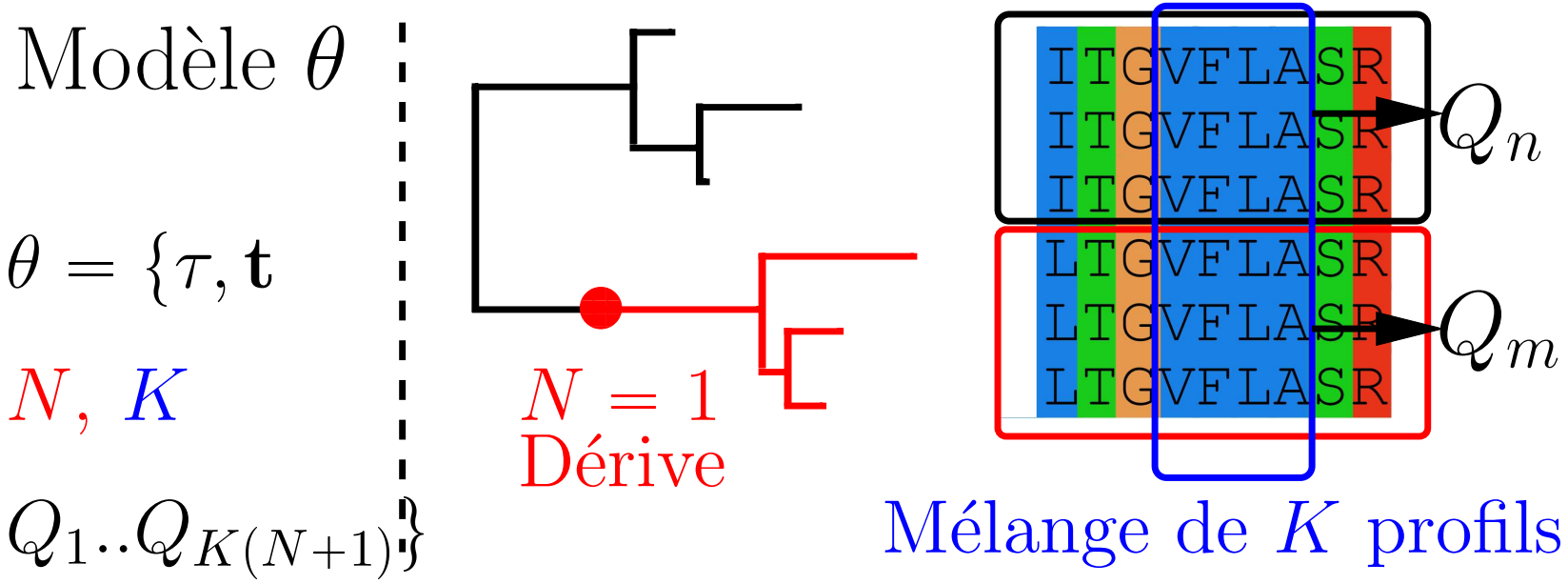
$Q_1..Q_{N+1}\}$



- 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).

Contribution 2: Relaxer l'hypothèse d'homogénéité



- 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.**

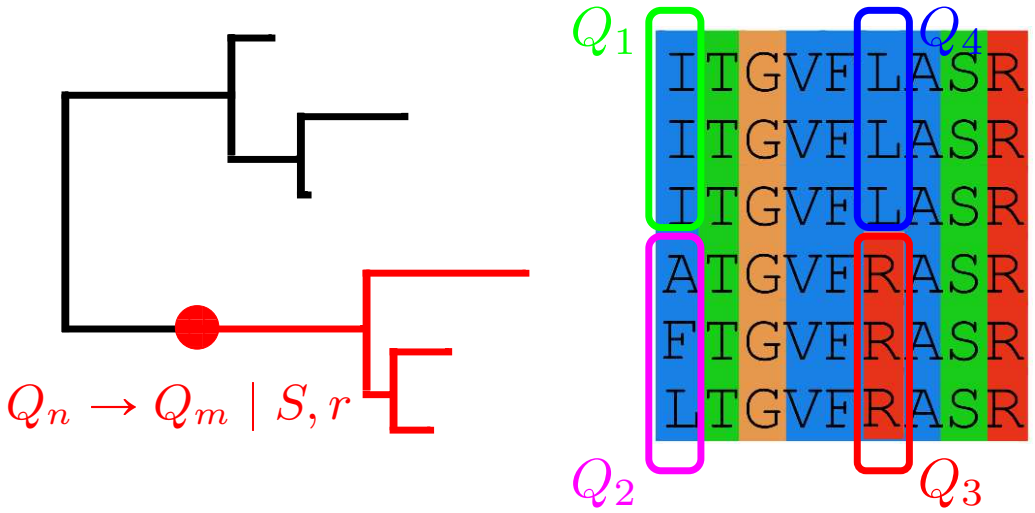
A site- and time-heterogeneous model of amino-acid replacement.

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

Contribution 3: Relaxer l'hypothèse d'homogénéité

Modèle θ

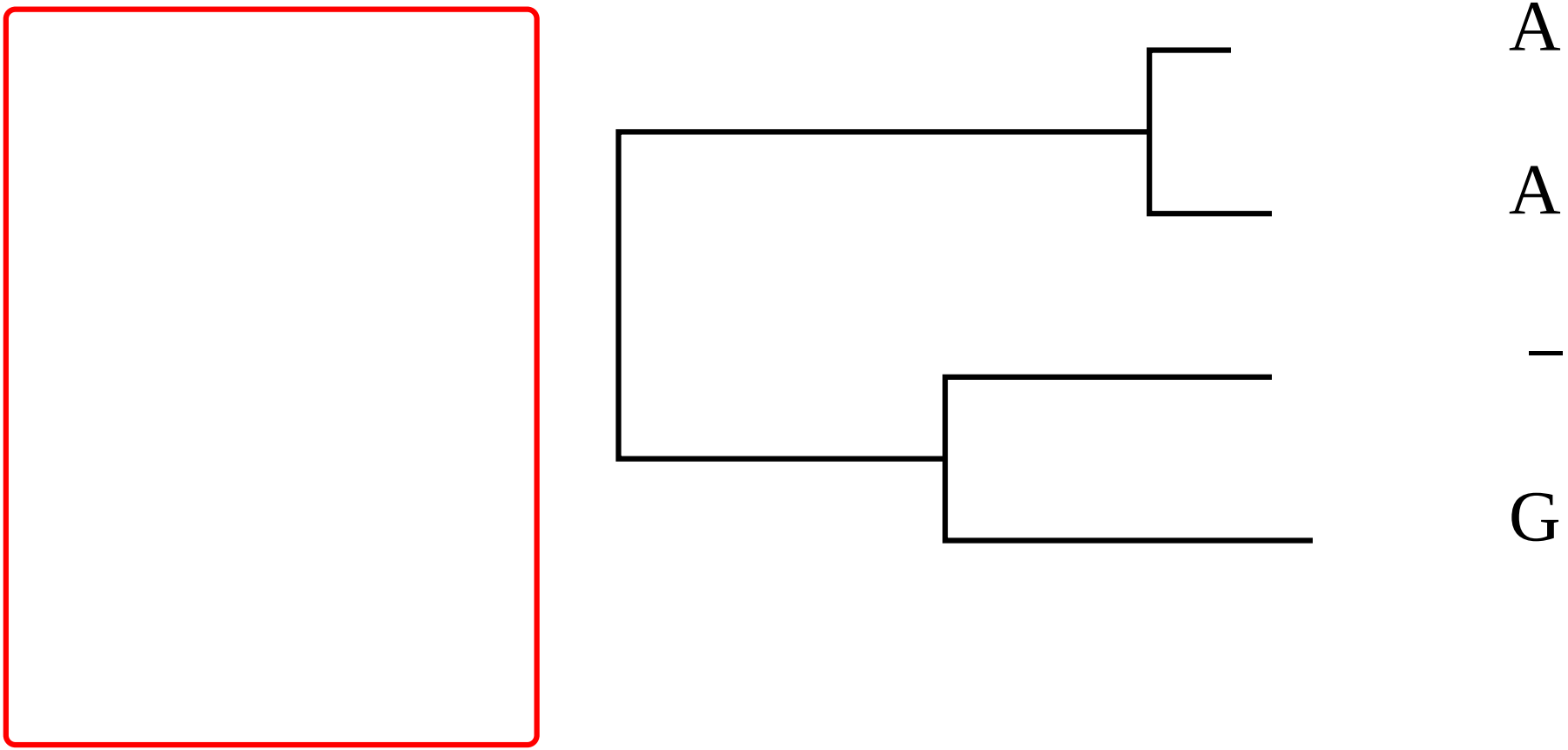
$\theta = \{\tau, \mathbf{t}$
N, K
$Q_1..Q_{K\times N},$
$S, r\}$



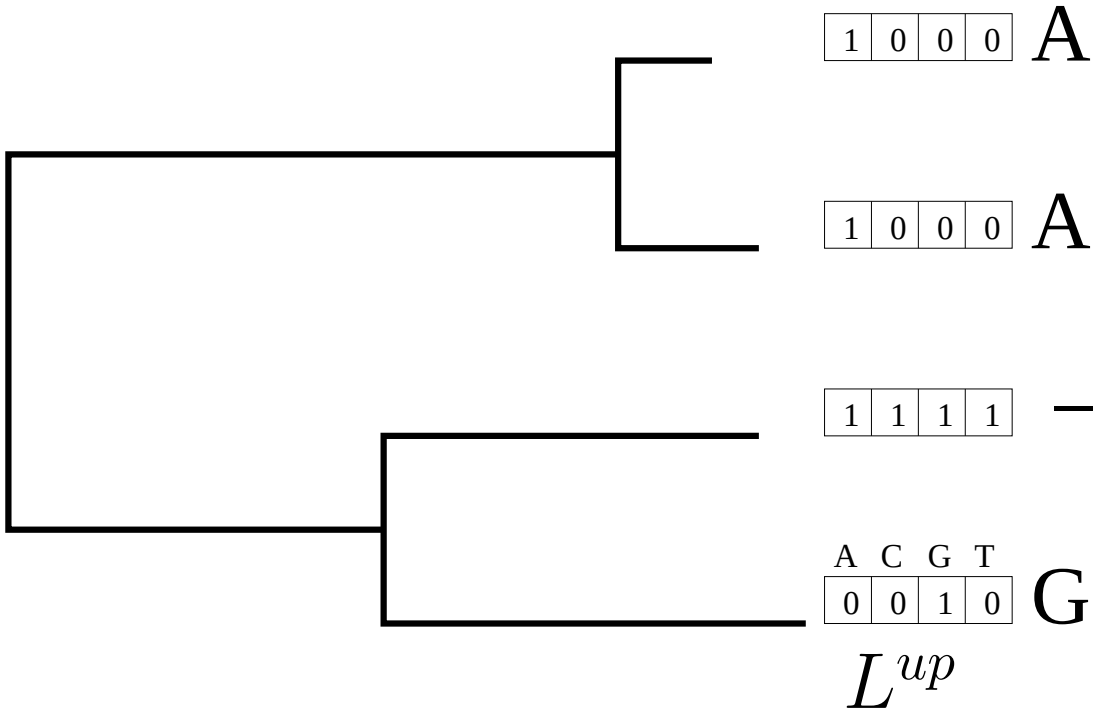
Variations spécifiques aux sites

- Modélisation de <u>N taux</u> (ex: Q_1: lent, Q_2: rapide) ET
- de <u>K profils biochimiques</u> (ex: Q_3: hydrophile, Q_4: hydrophobe).
- <u>$K \times N$</u> processus $\mathbf{Q}$ et un processus additionnel S de taux r sont combinés en un processus de Markov Markov-Modulé: $\mathcal{M} = (I \oslash \mathbf{Q}) + (r \times S \otimes I)$.

Calcul récursif de la vraisemblance (Felsenstein 1981)

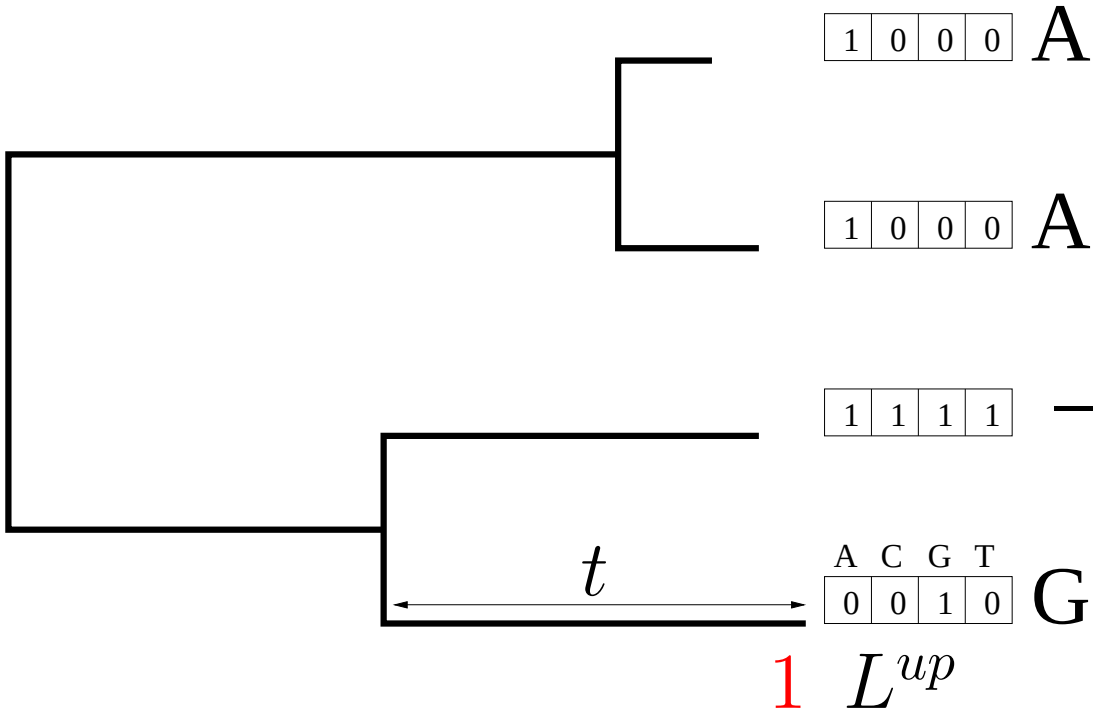


Calcul récursif de la vraisemblance (Felsenstein 1981)



Calcul récursif de la vraisemblance (Felsenstein 1981)

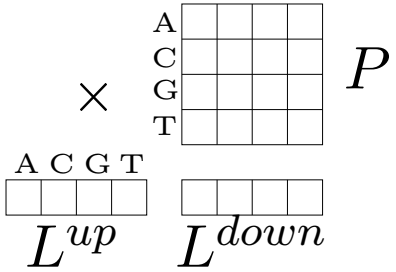
1) $P = e^{tQ}$

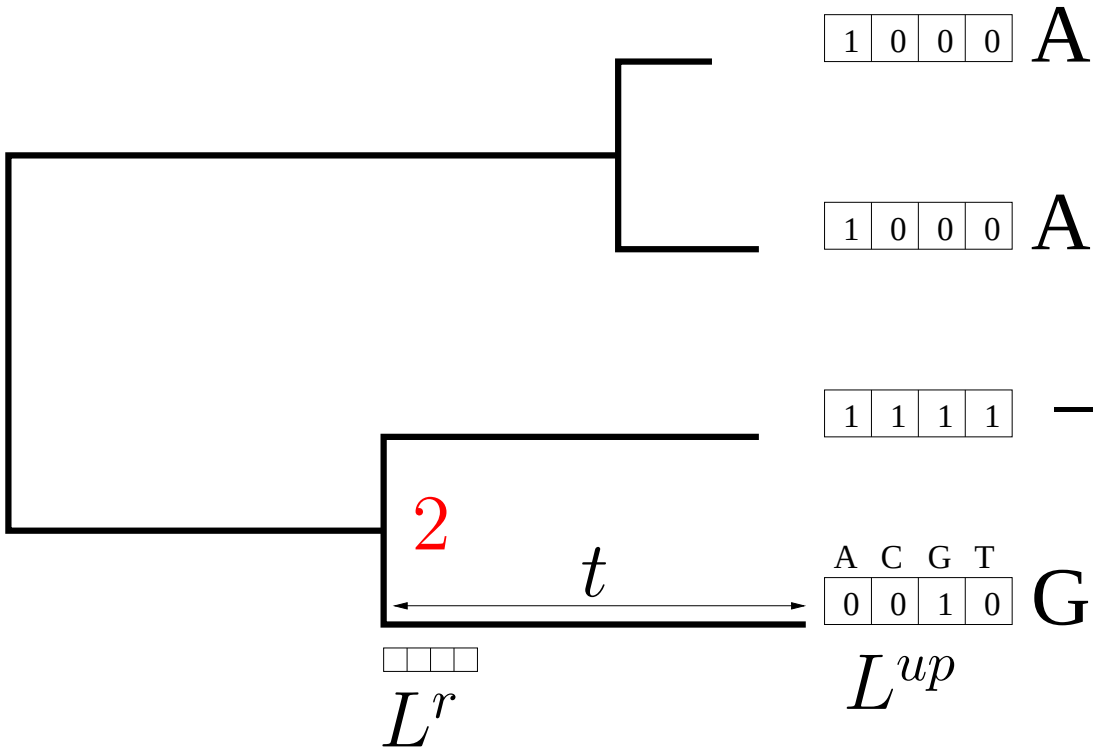


Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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

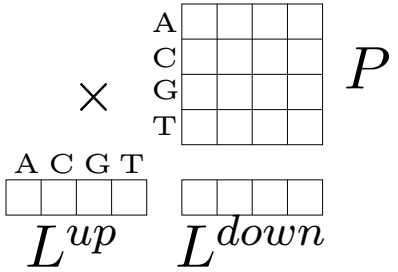


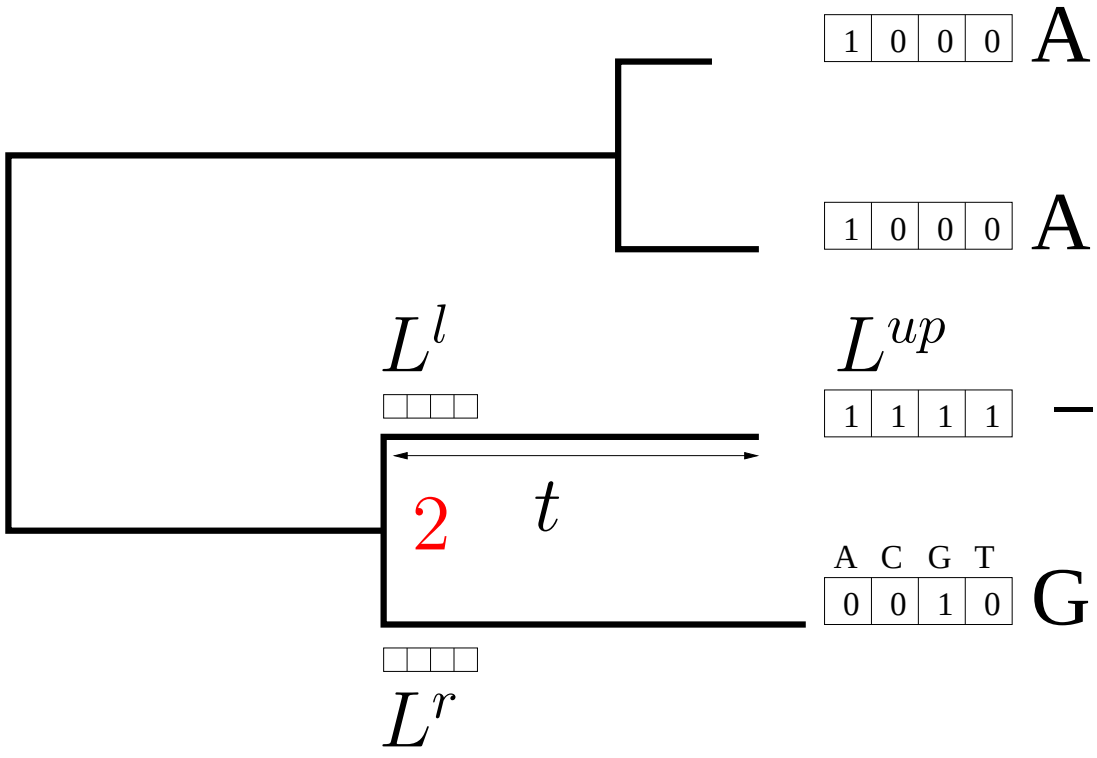


Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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

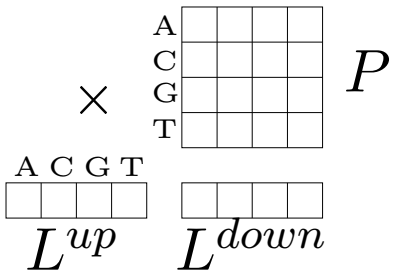




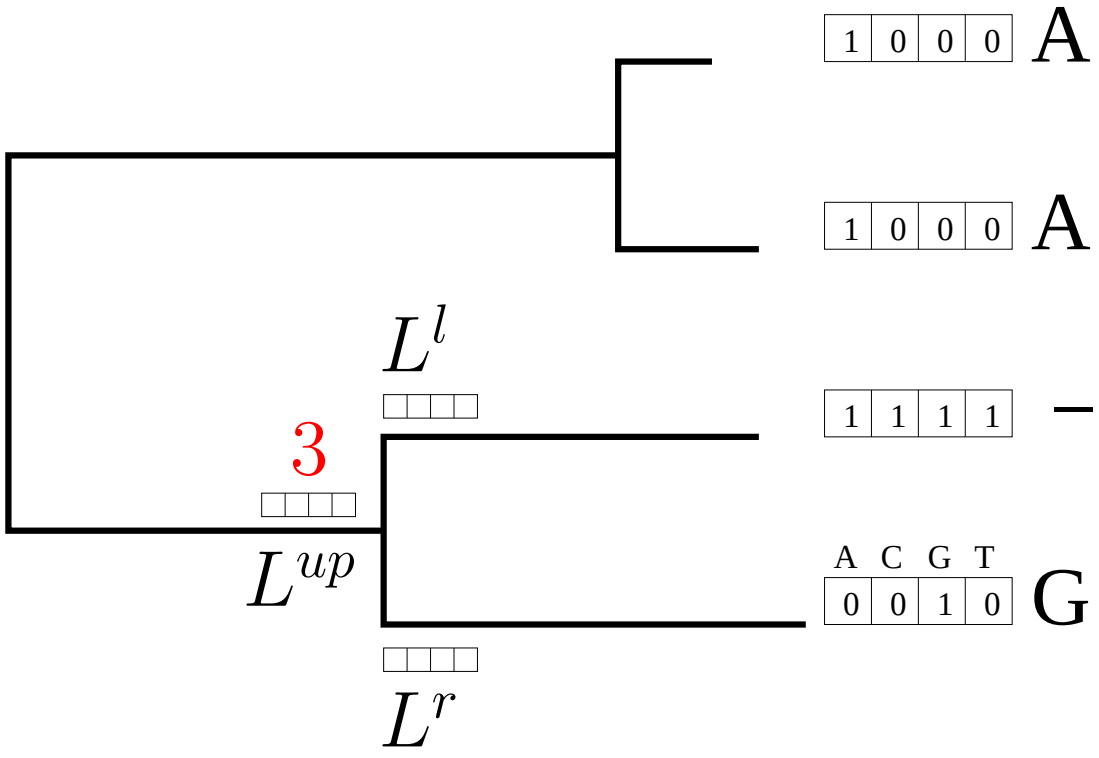
Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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



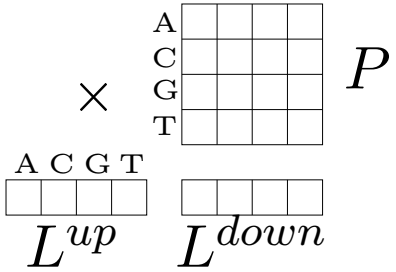
3) $L_i^{up} = L_i^l L_i^r$



Calcul récursif de la vraisemblance (Felsenstein 1981)

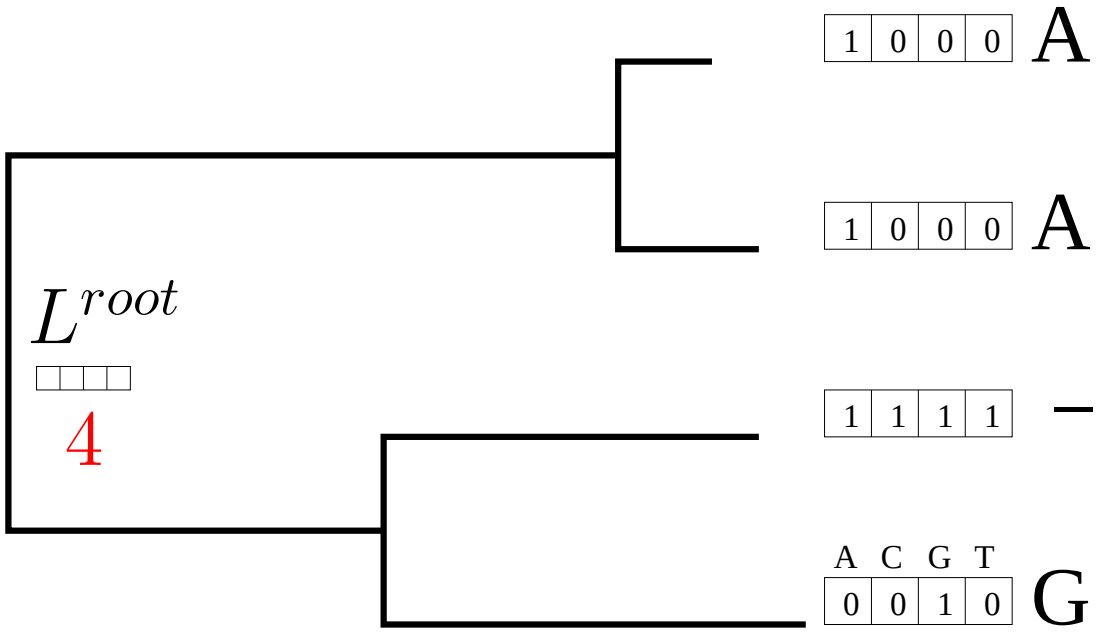
1) $P = e^{tQ}$

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



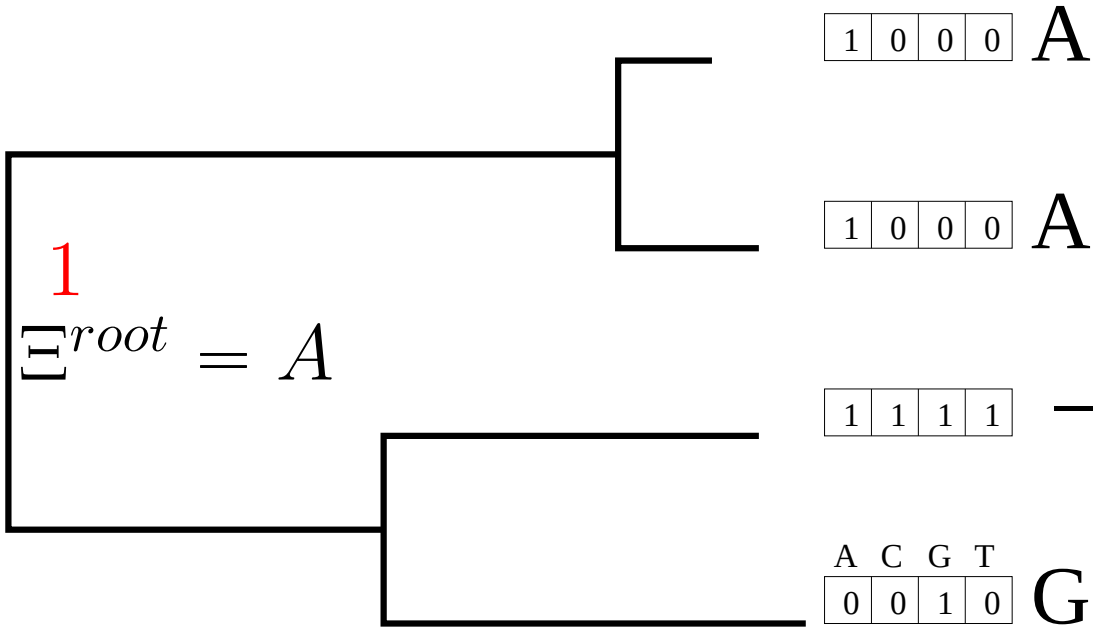
3) $L_i^{up} = L_i^l L_i^r$

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



Construire les séquences ancestrales

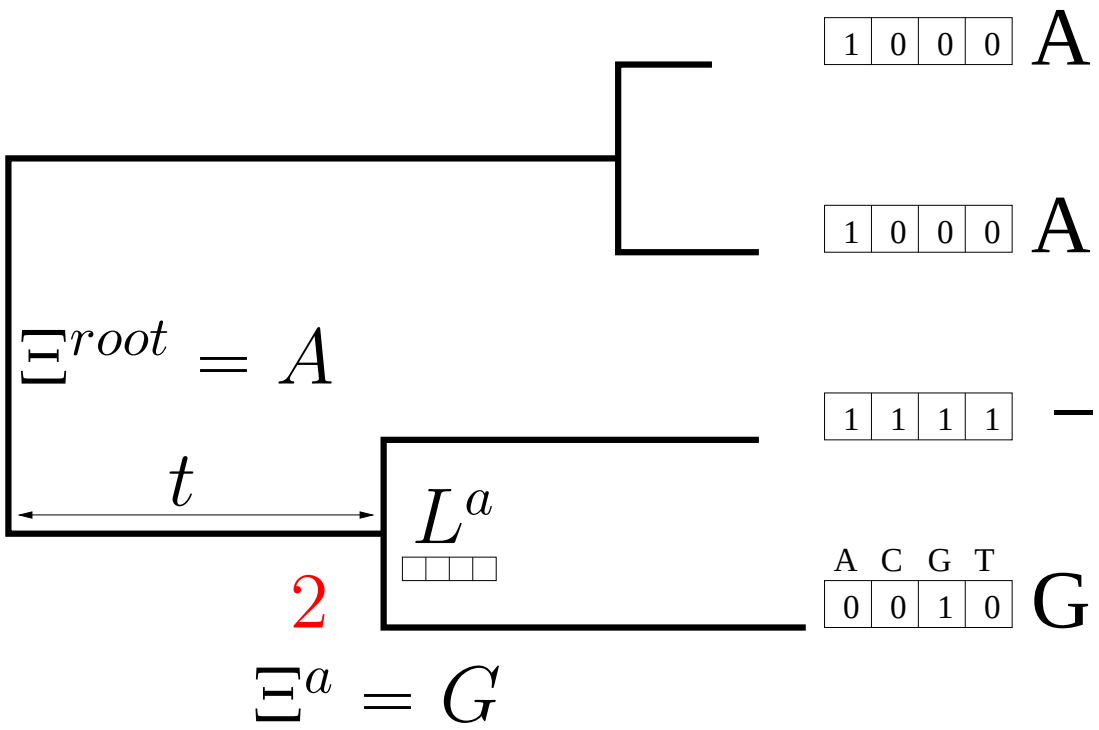
1) $\Xi^{root} \sim \pi_i L_i^{root}$



Construire les séquences ancestrales

1) $\Xi^{root} \sim \pi_i L_i^{root}$

2) $\Xi^{a} \sim P_{\Xi^{root},i} L_i^{a}$

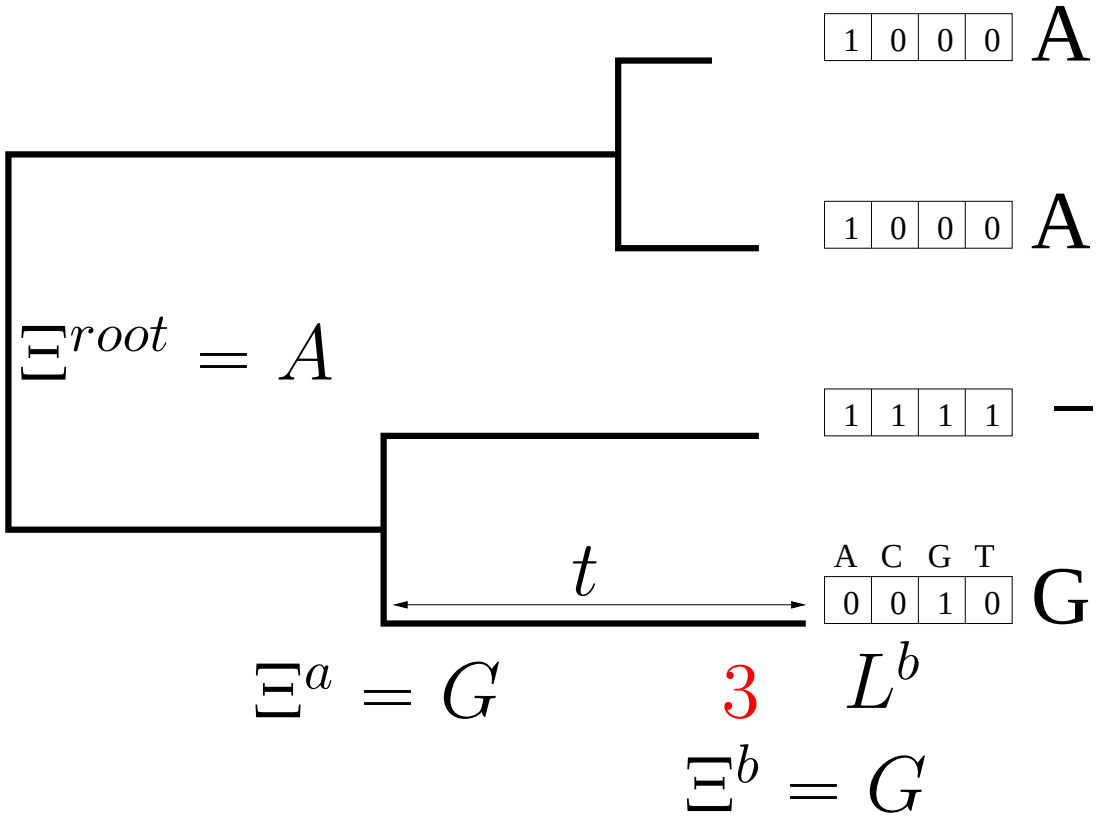


Construire les séquences ancestrales

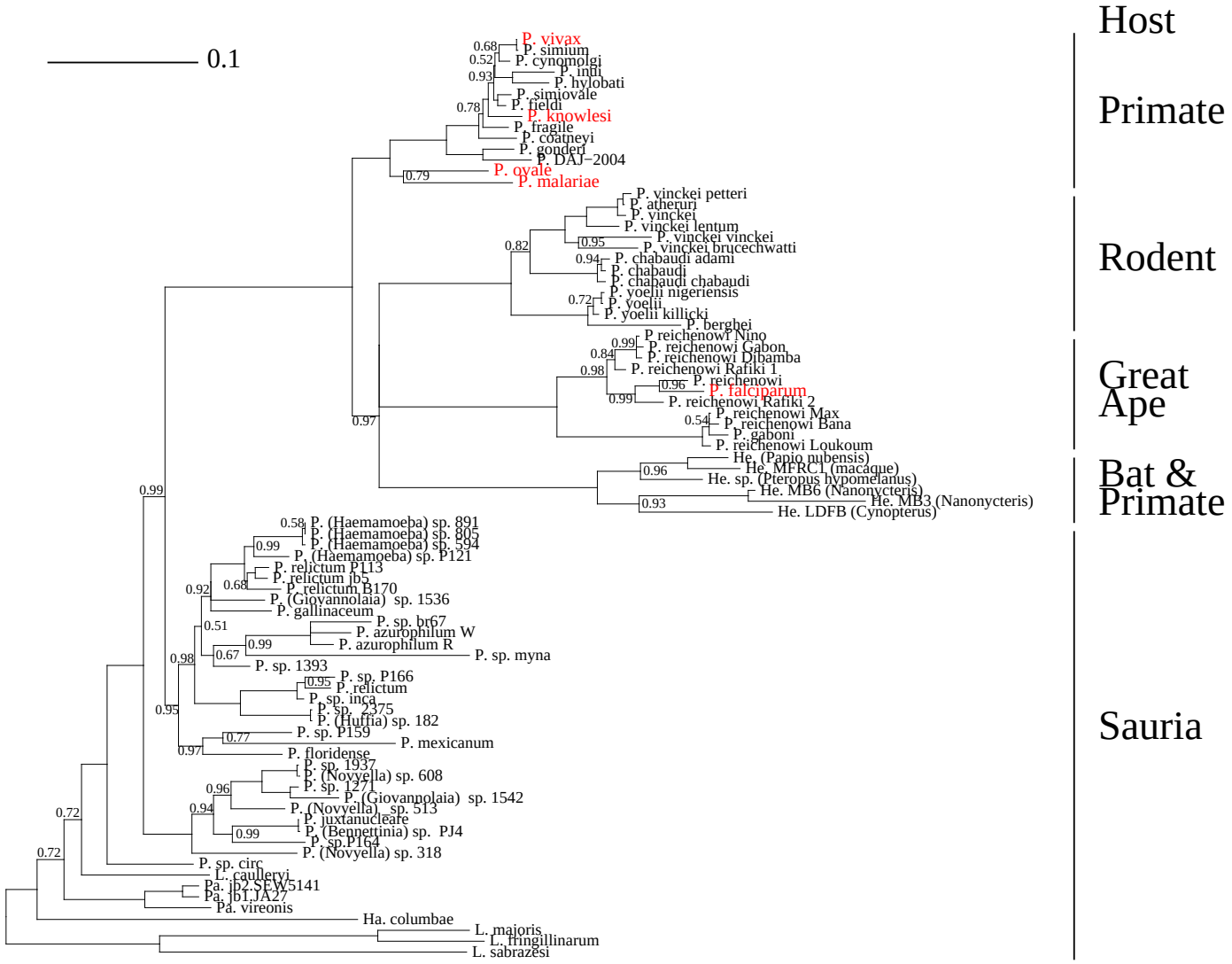
1) $\Xi^{root} \sim \pi_i L_i^{root}$

2) $\Xi^{a} \sim P_{\Xi^{root},i} L_i^{a}$

3) $\Xi^{b} \sim P_{\Xi^{a},i} L_i^{b}$



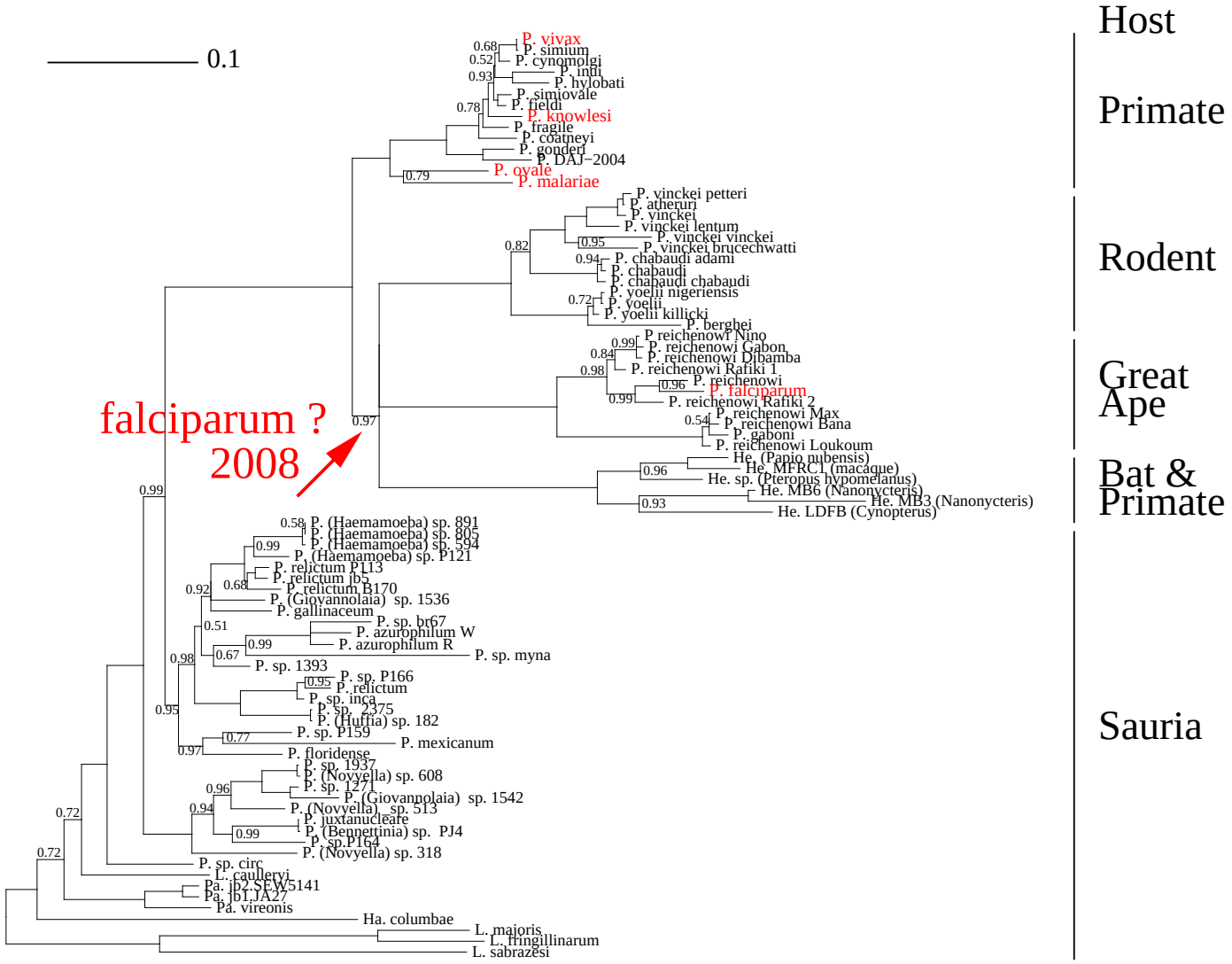
Contribution 4: Phylogénie des parasites malariaux



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).

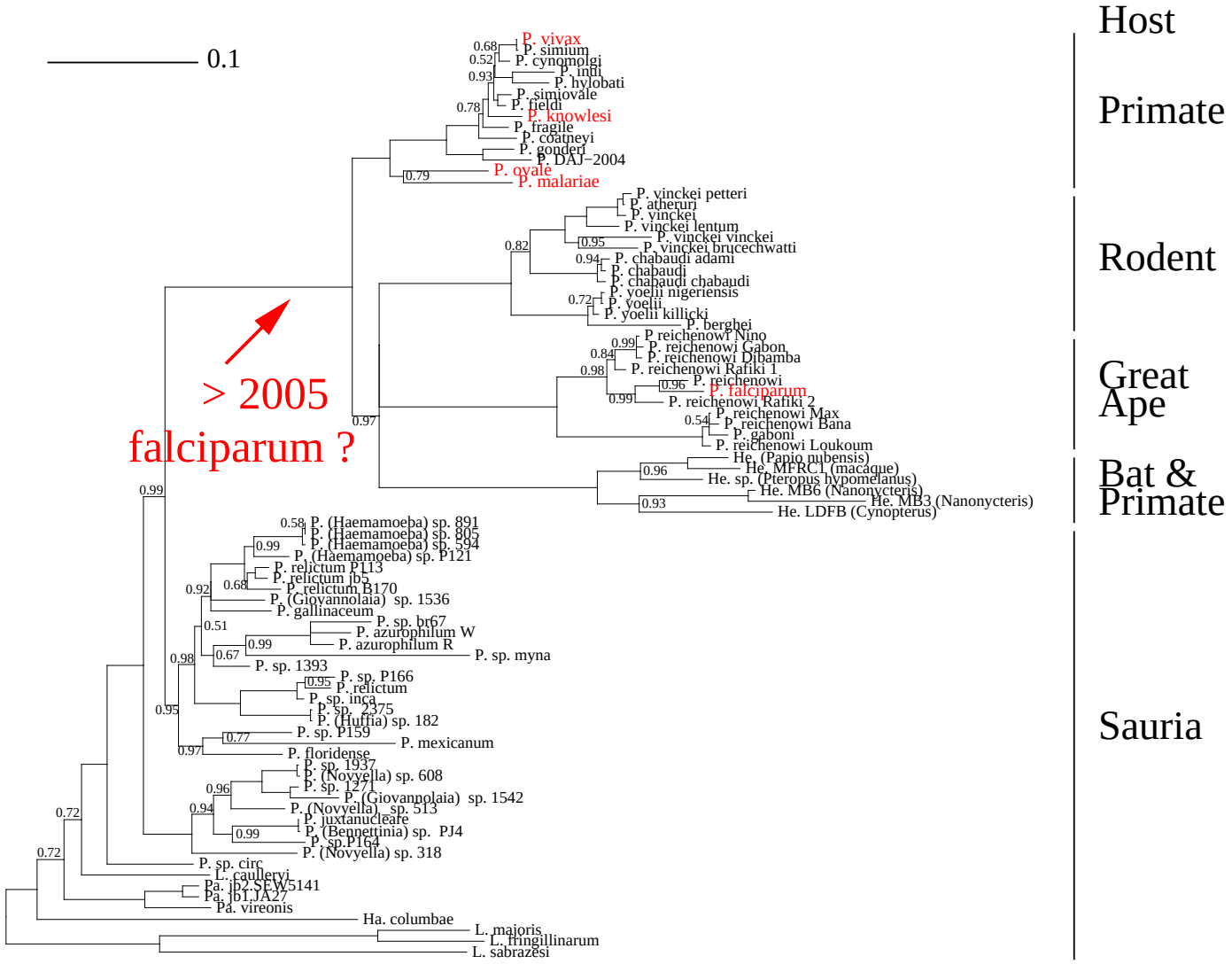
Contribution 4: Phylogénie des parasites malariaux



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).

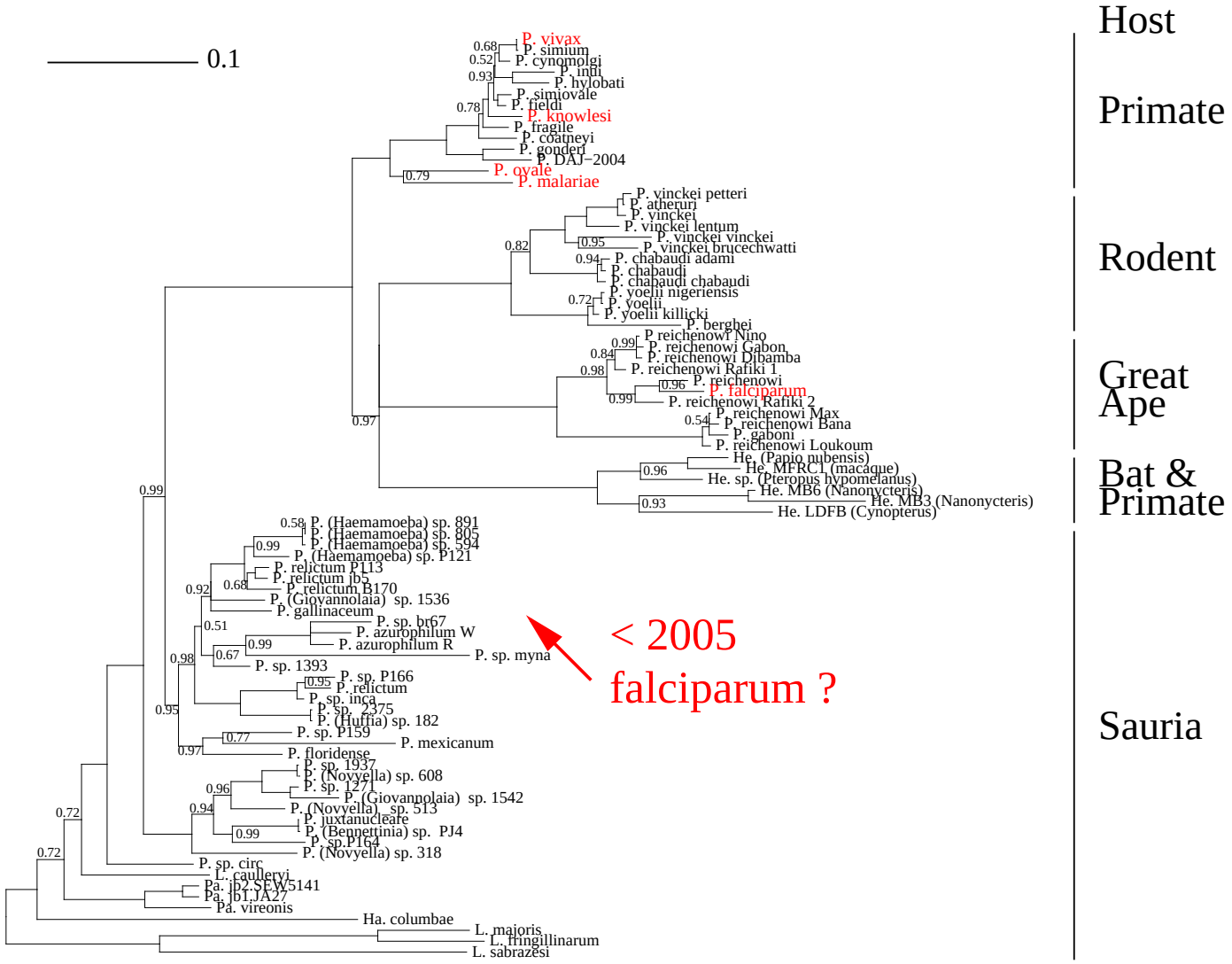
Contribution 4: Phylogénie des parasites malariaux



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).

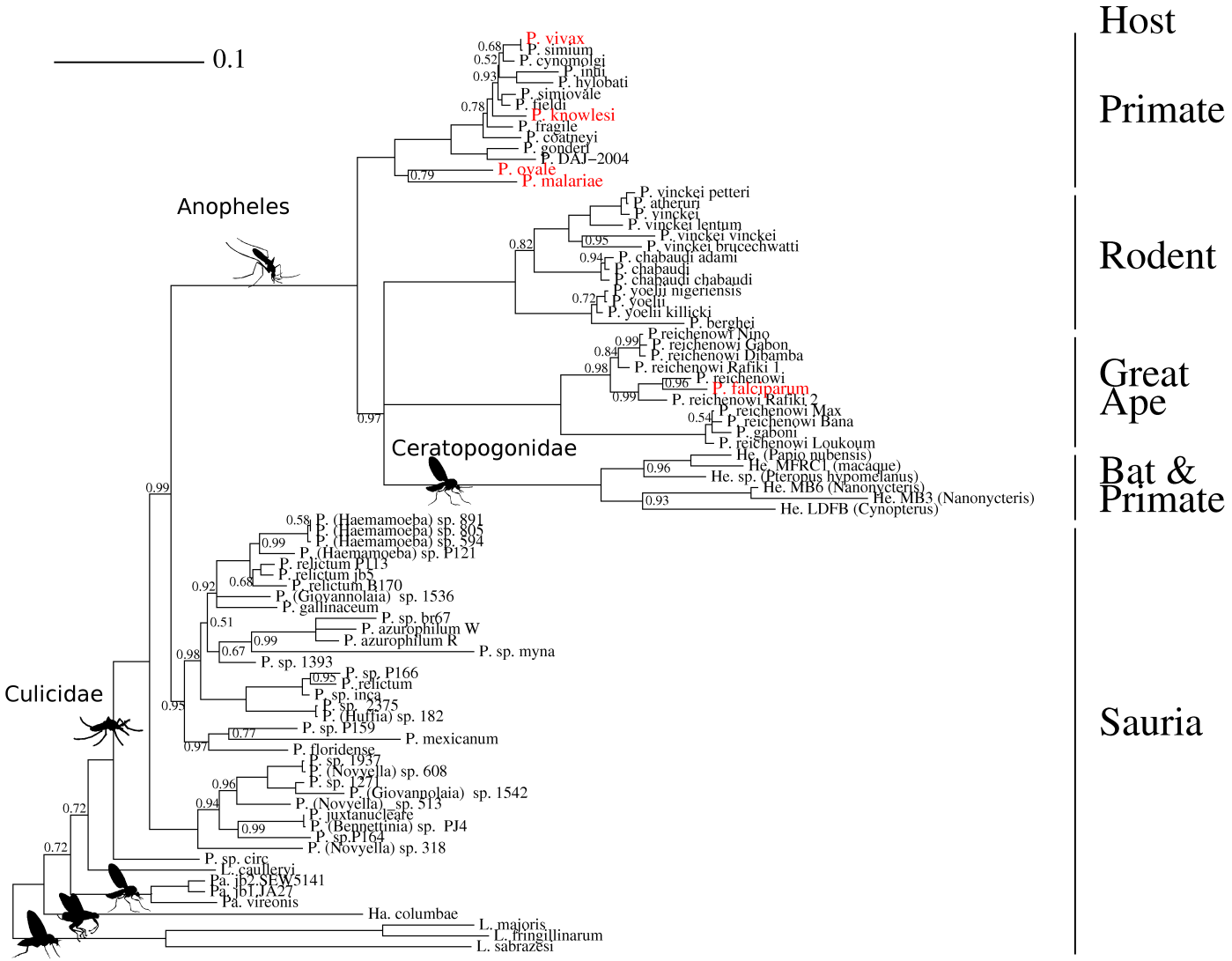
Contribution 4: Phylogénie des parasites malariaux



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).

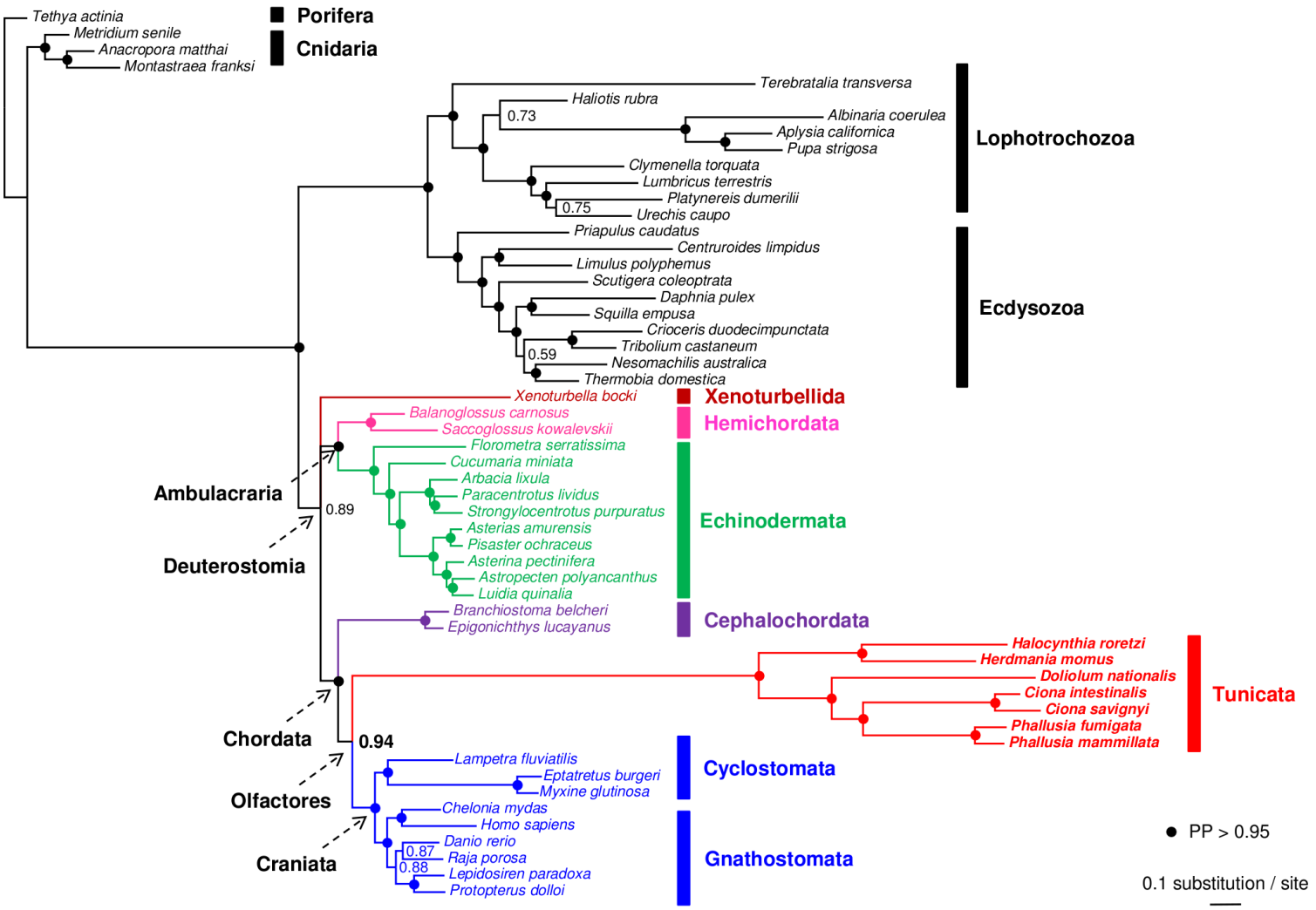
Contribution 4: Phylogénie des parasites malariaux



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).

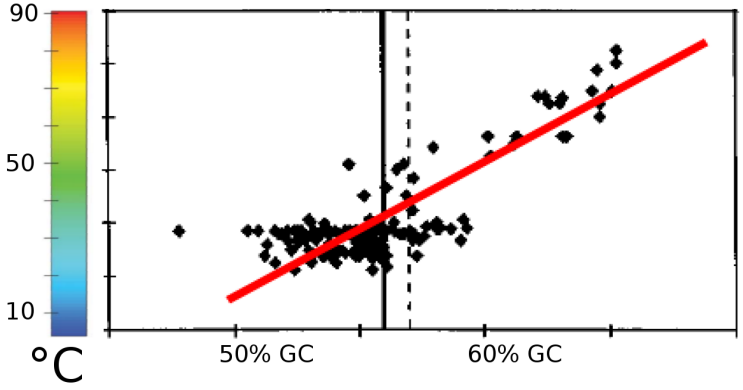
Contribution 5: Phylogénie des métazoaires



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).

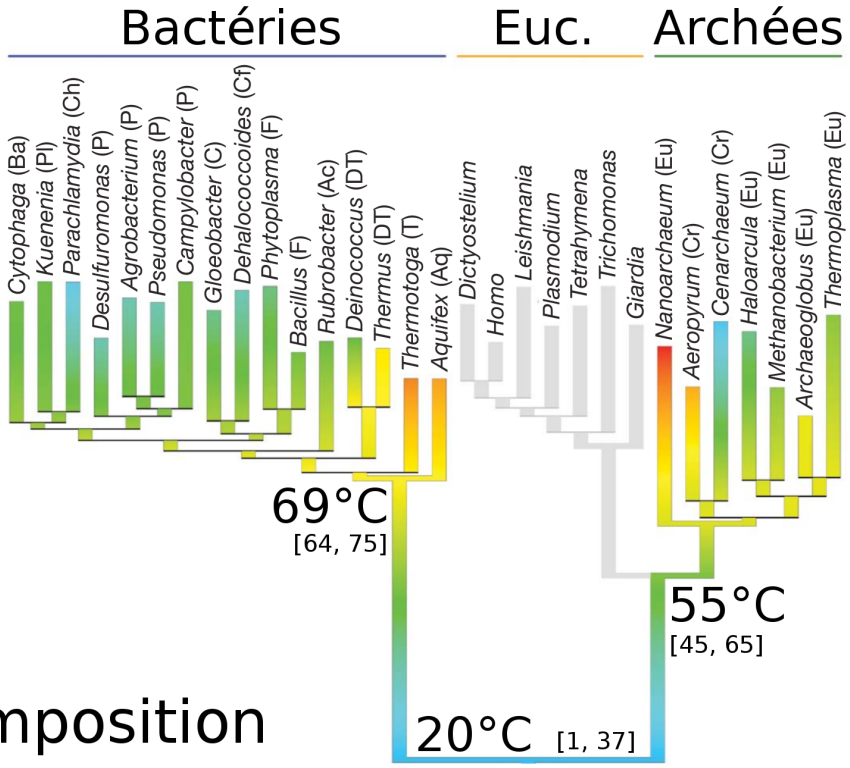
Contribution 6: Températures des paléo-environnements

ARN ribosomiques



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

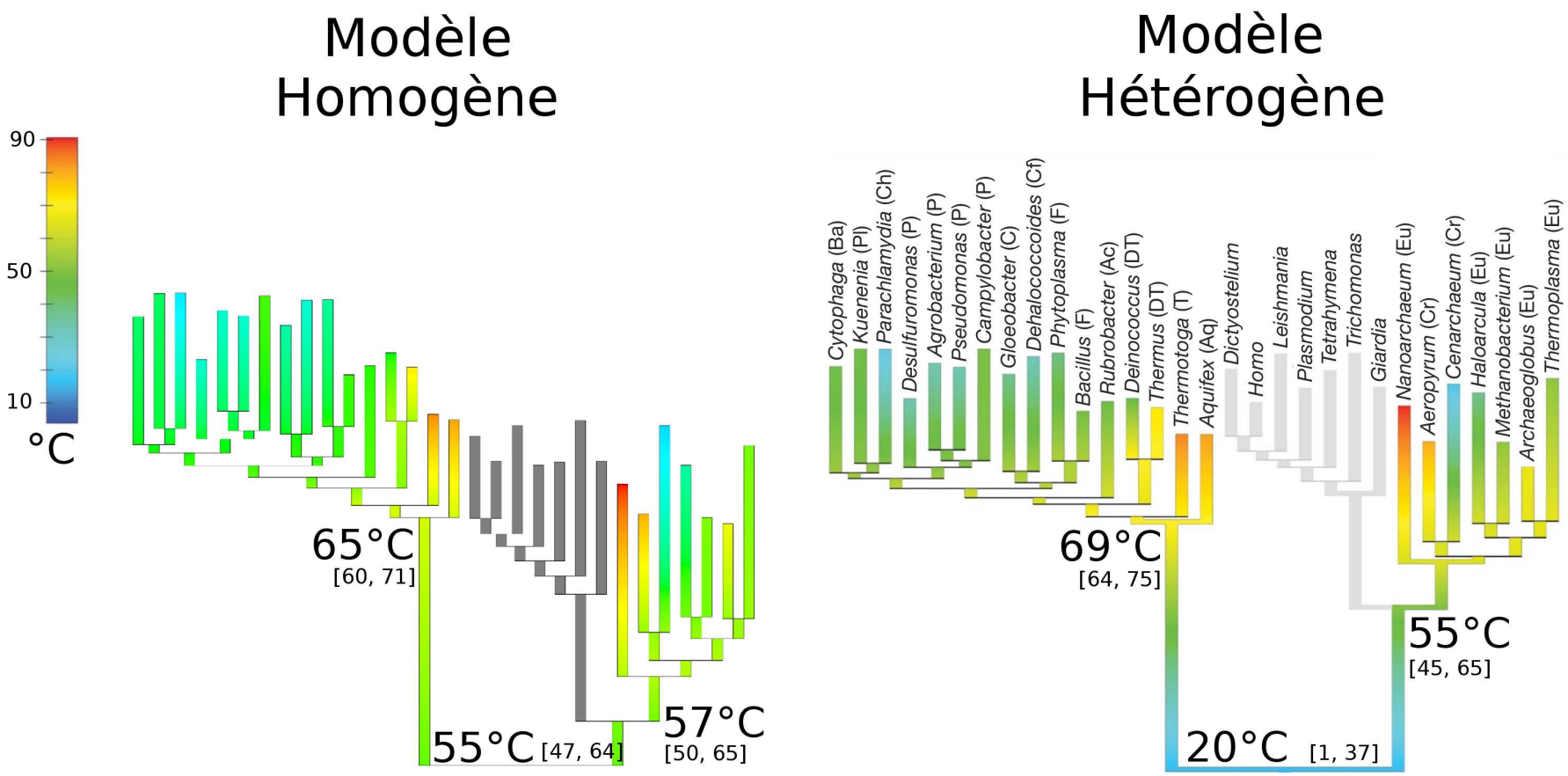
Protéines



Corrélation température/composition
et inférence des compositions ancestrales

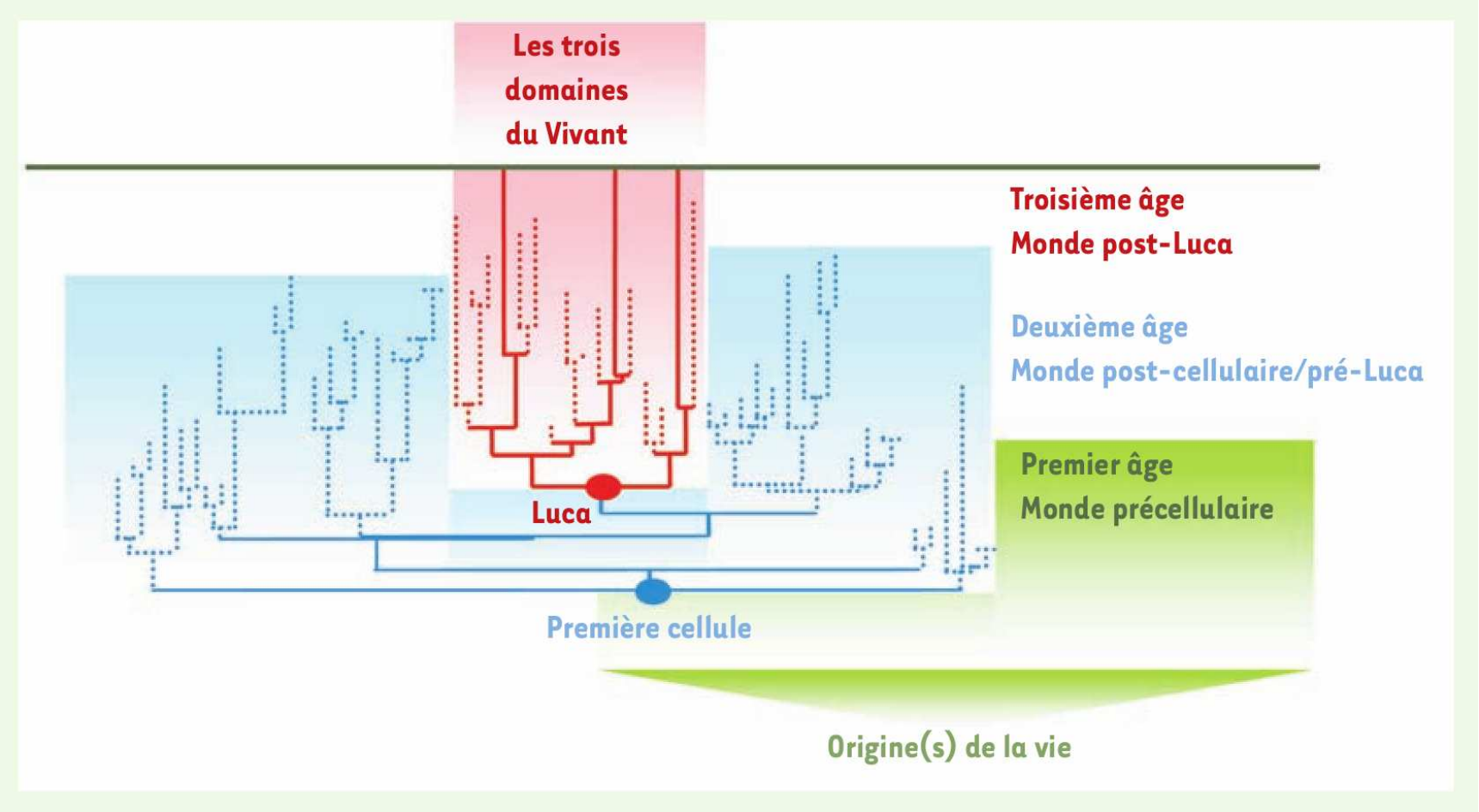
Parallel Adaptations to High Temperatures in the Archean Eon.
Boussau*, **Blanquart***, Necsulea, Lartillot, & Gouy. Nature (2008)
(* co premier auteur).

Contribution 6: Températures des paléo-environnements



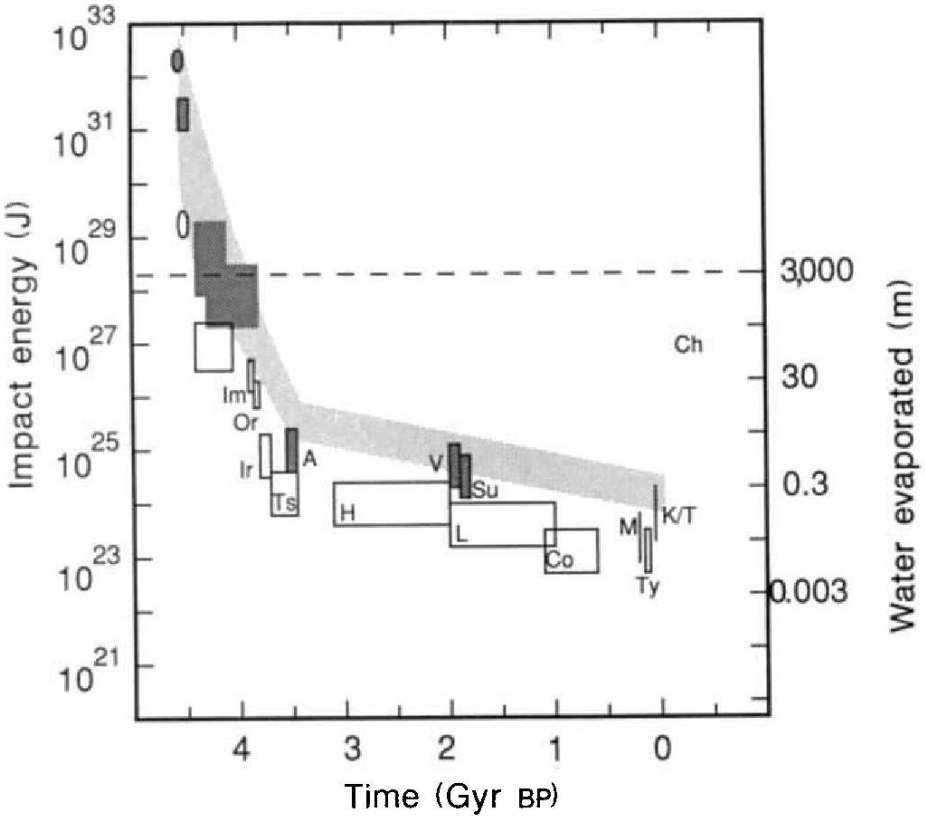
Parallel Adaptations to High Temperatures in the Archean Eon.
Boussau*, **Blanquart***, Necsulea, Lartillot, & Gouy. Nature (2008)
(* co premier auteur).

Ecologie des paléo-environnements Archéen et Hadéen



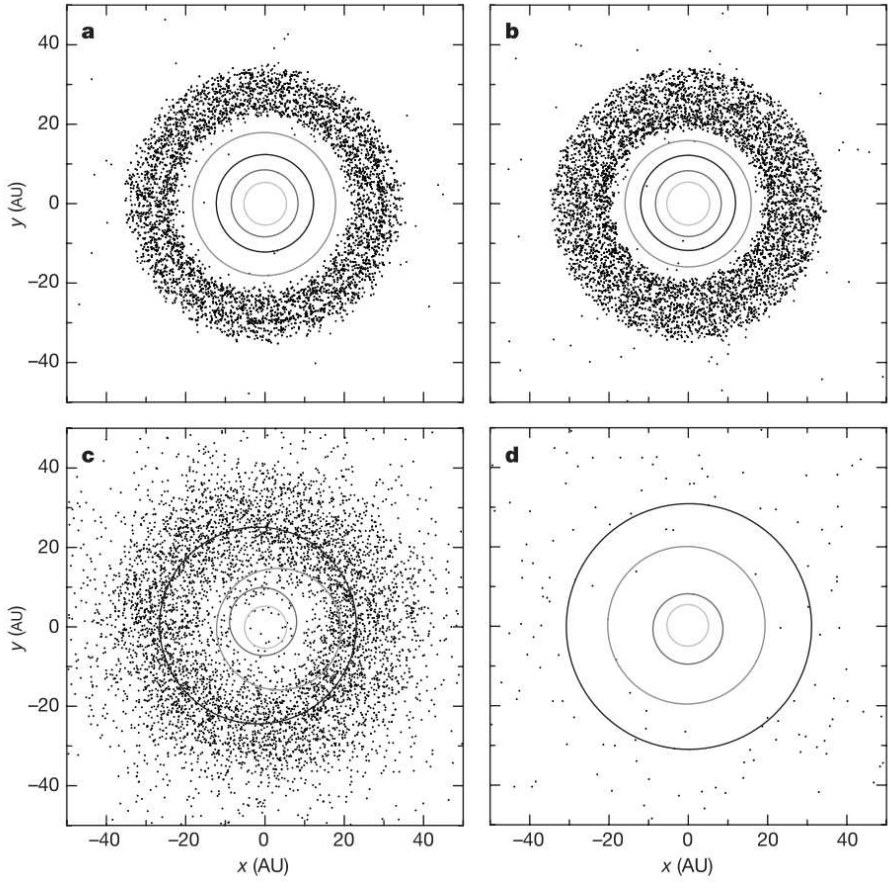
Luca : à la recherche du plus proche ancêtre commun universel.
Forterre, Gribaldo & Brochier. Médecine/Science (2005).

Environnement inter-planétaire



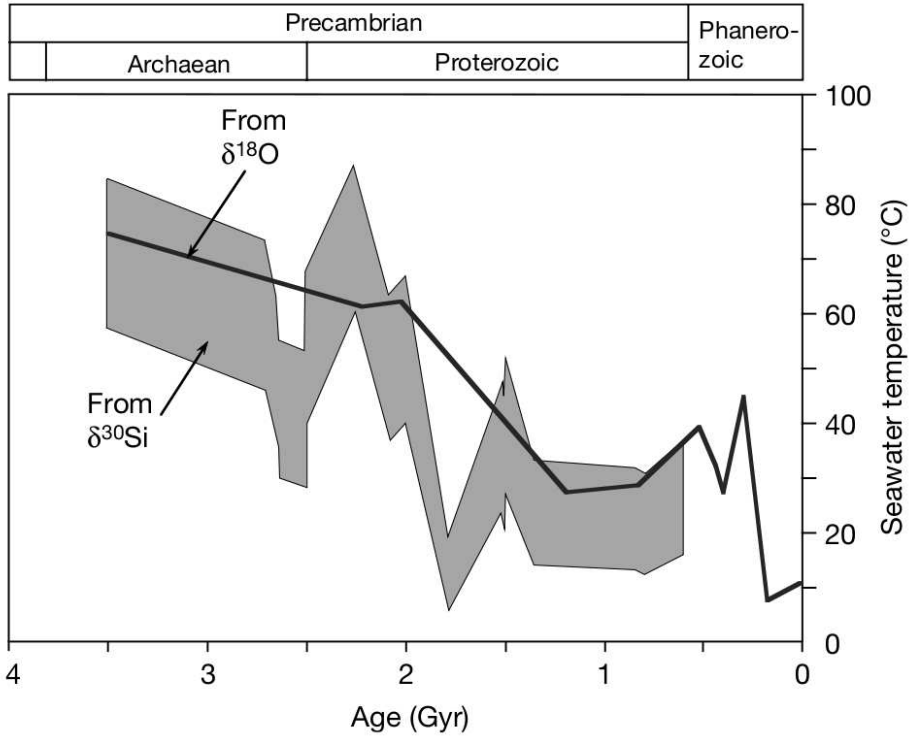
Annihilation of ecosystems by large asteroid impacts on the early Earth. Sleep, Zahnle, Kasting & Morowitz. Nature (1989).

Le dernier bombardement intense, -3.7 Gy



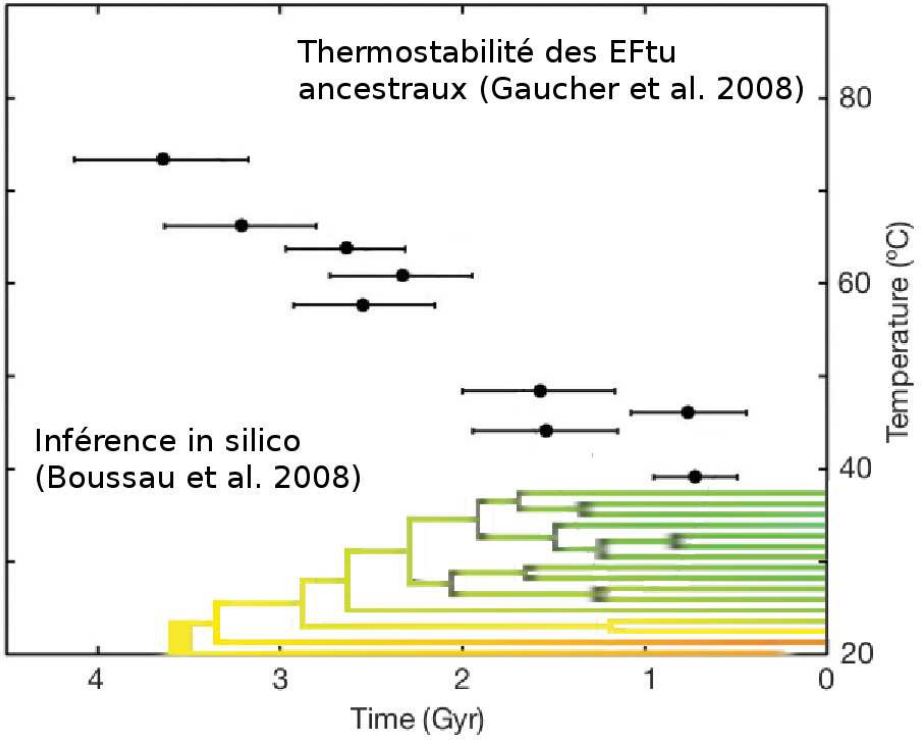
Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Gomes, Levison, Tsiganis & Morbidelli. Nature (2005).

Température des paléo-océans



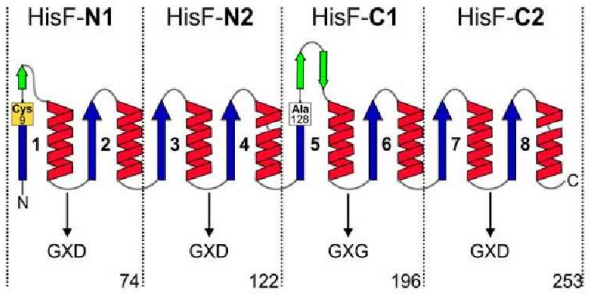
Palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. Robert & Chaussidon. Nature (2006).

Evolution de la thermophilie bactérienne



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

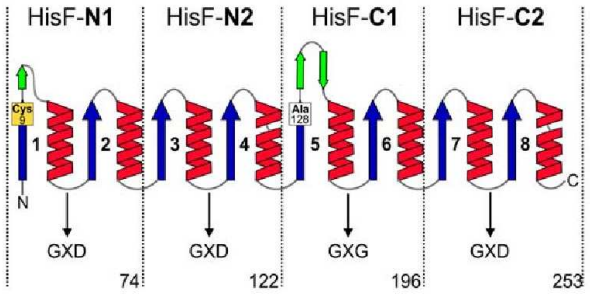
Contribution 7: Résurrection de gènes ancestraux



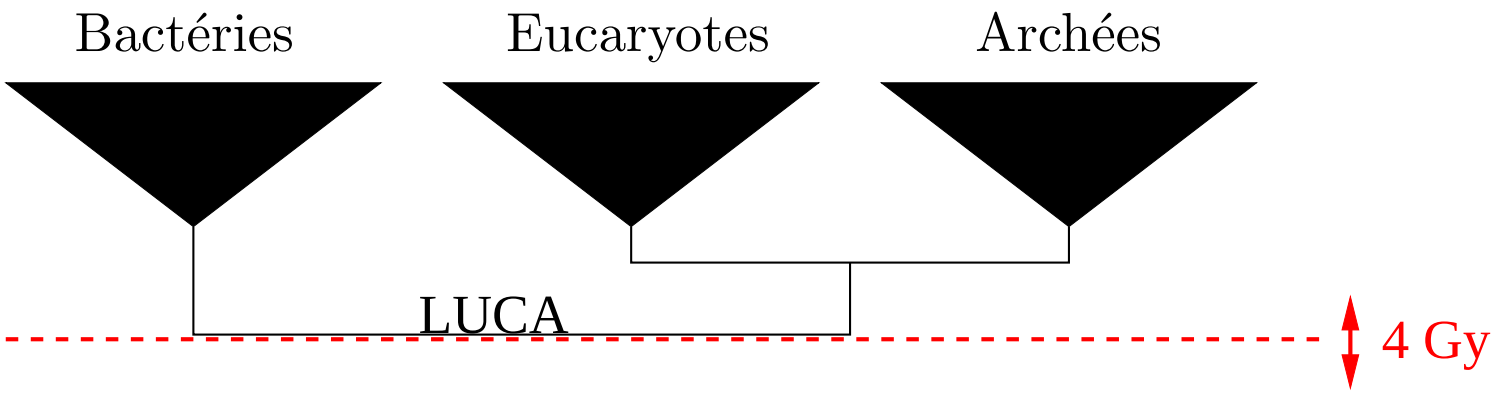
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA

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

Contribution 7: Résurrection de gènes ancestraux



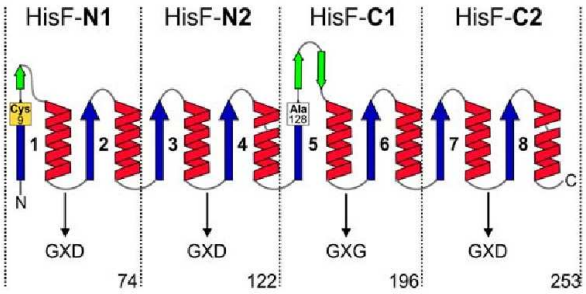
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA



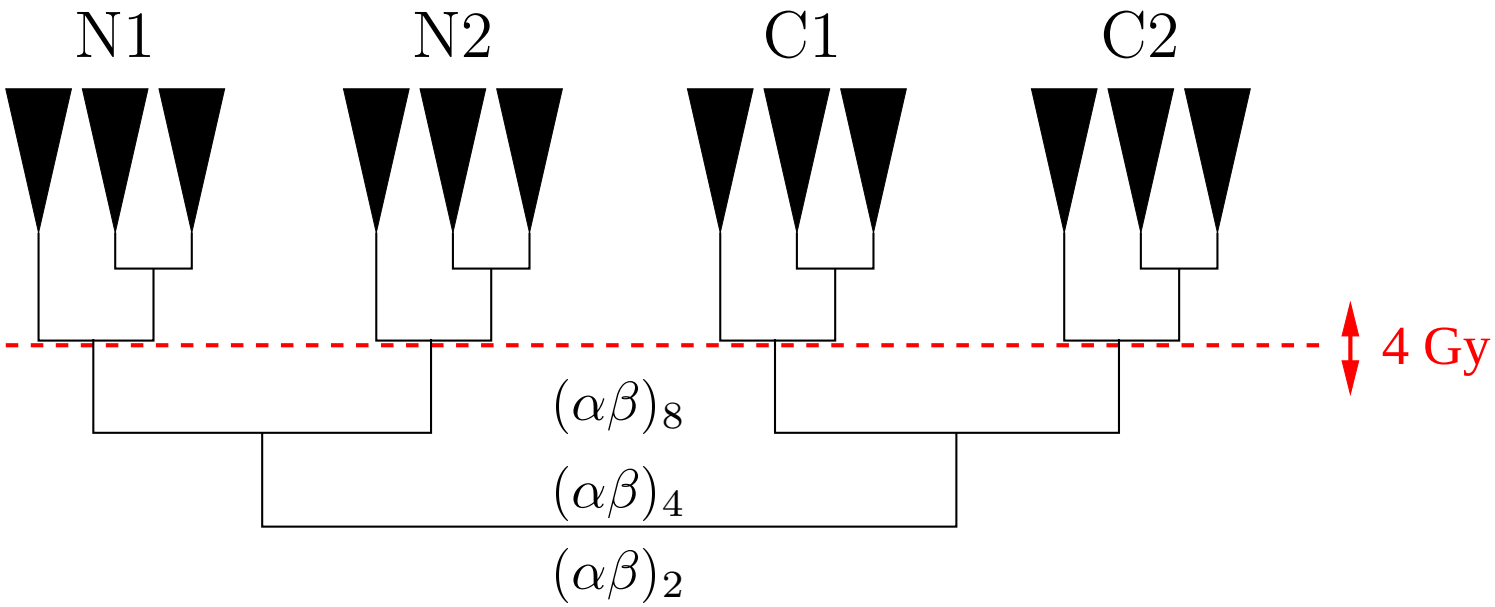
Arbre shématique des 8-mères $(\alpha\beta)_8$

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

Contribution 7: Résurrection de gènes ancestraux



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

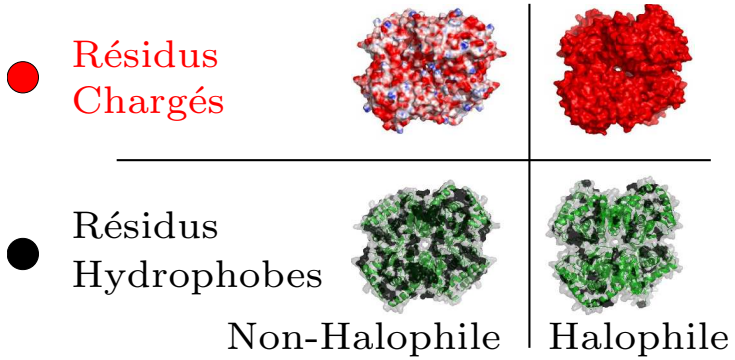


Arbre shématique des dimères $(\alpha\beta)_2$

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

Conclusion et perspectives

- Projet en cours: Paléo-biologie synthétique



Adaptations de Malates et de Lactates déshydrogénases bactériennes et archées aux conditions de vies halophiles et thermophiles.

- Optimisation des outils et modèles existants.
- Développement de méthodes d'analyse phylogénétique plus rapides et adaptées aux données issues de la méta-génomique.
- Histoire des réarrangements génomiques et production de jeux de données pour la phylogénomique.