Les modèles phylogénétiques comme machines à remonter le temps

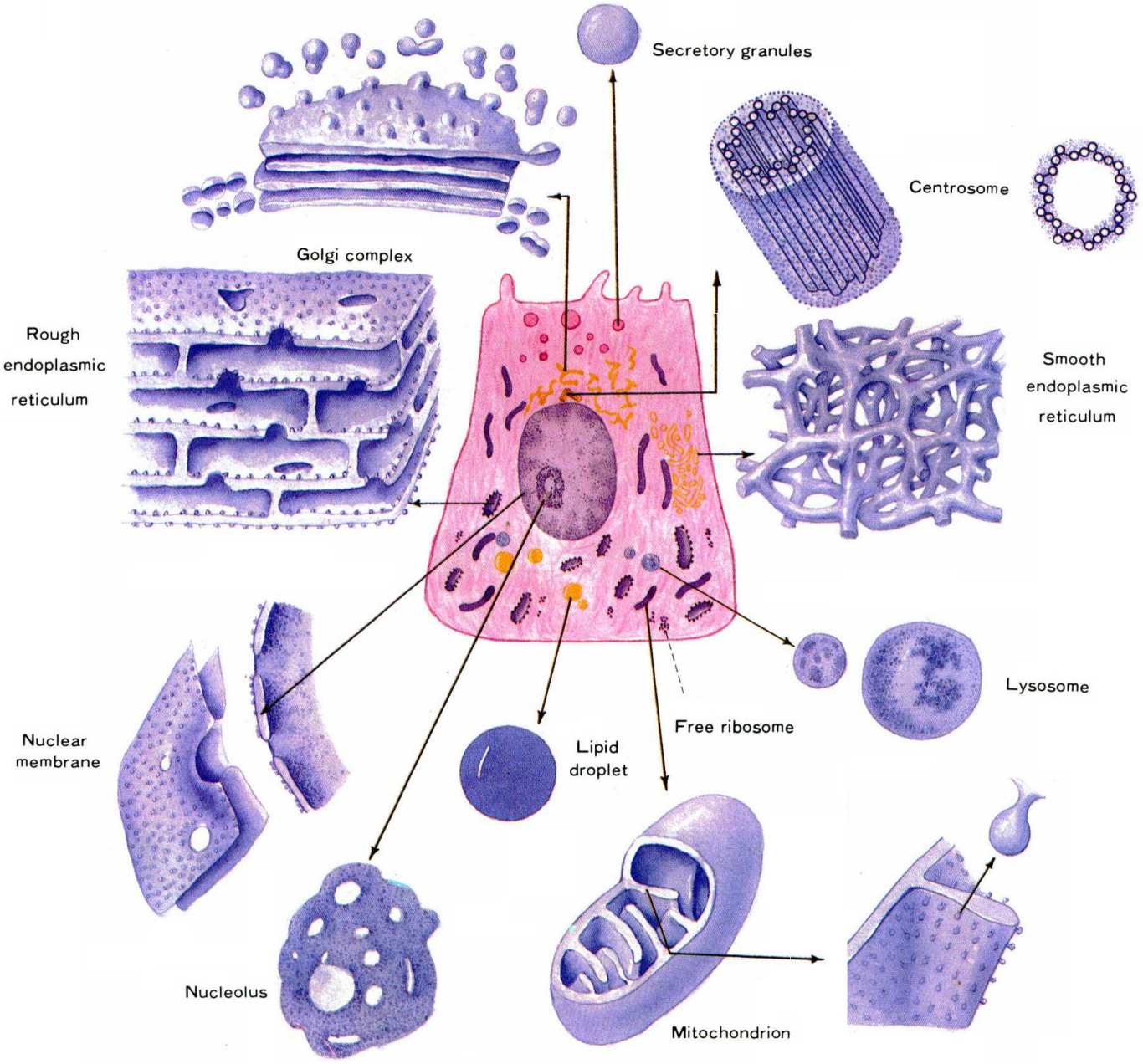
Samuel Blanquart, CR2 INRIA, EPI Bonsai

17 octobre 2011

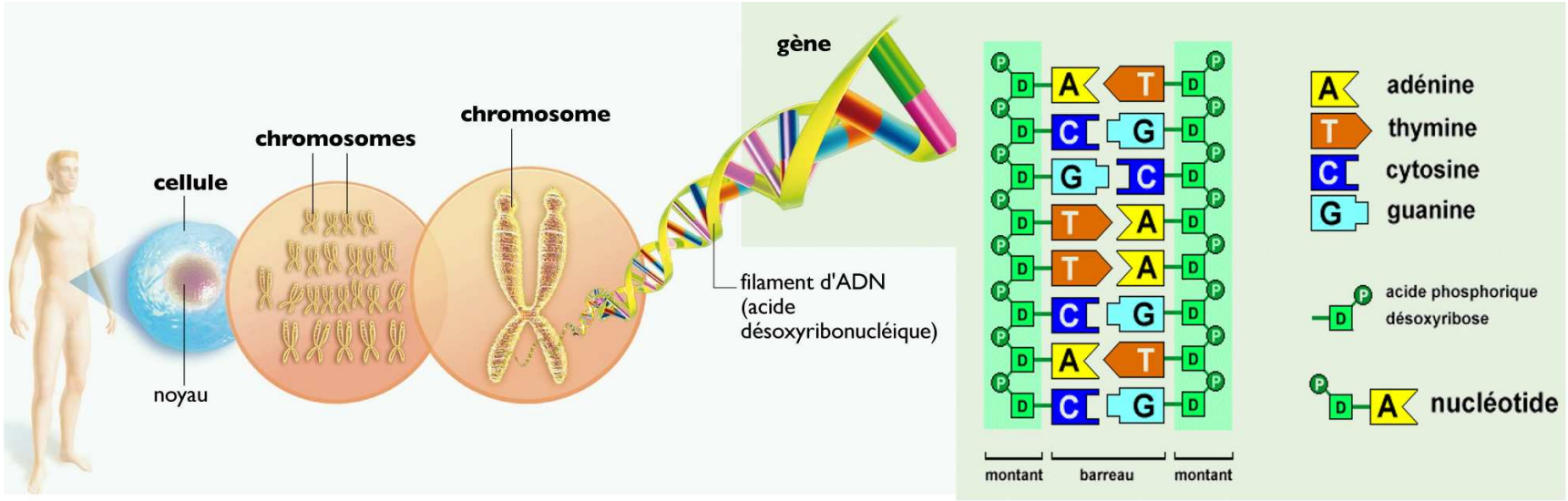
Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution.
Theodosius Dobzhansky.

Les organismes unicellulaires

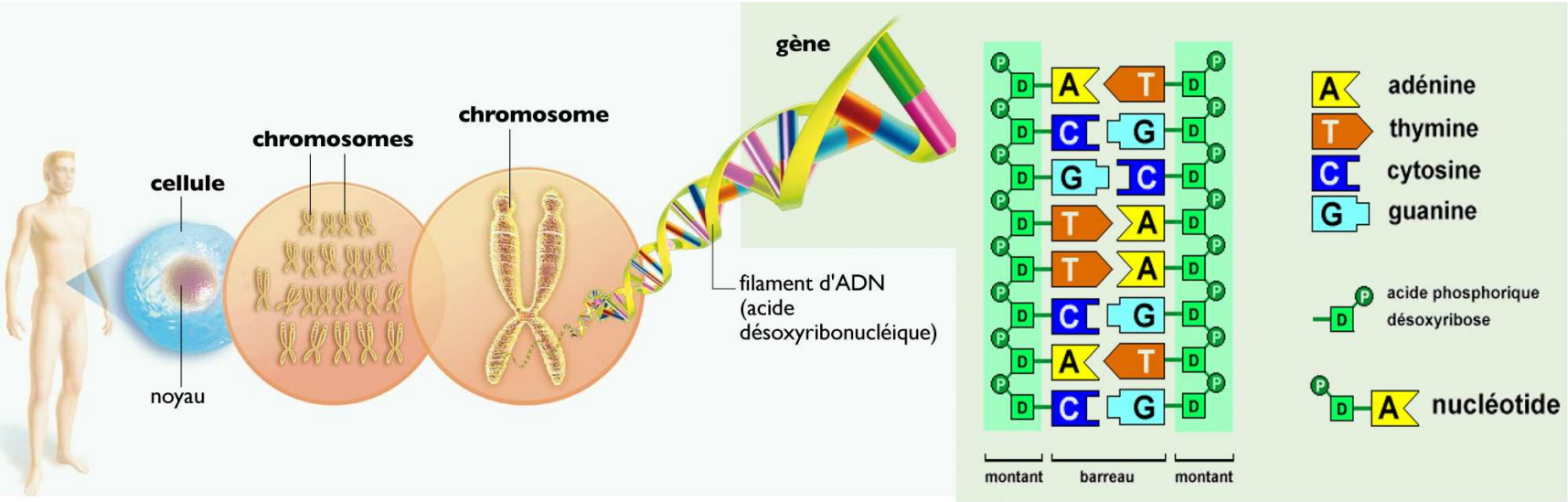
Les organes de la cellule

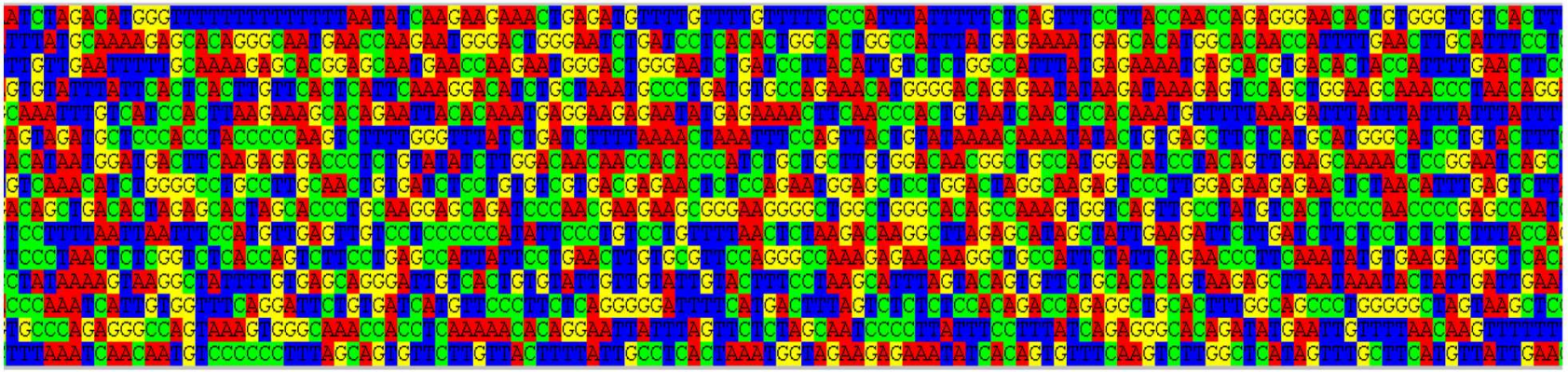


De l'organisme à l'information génétique

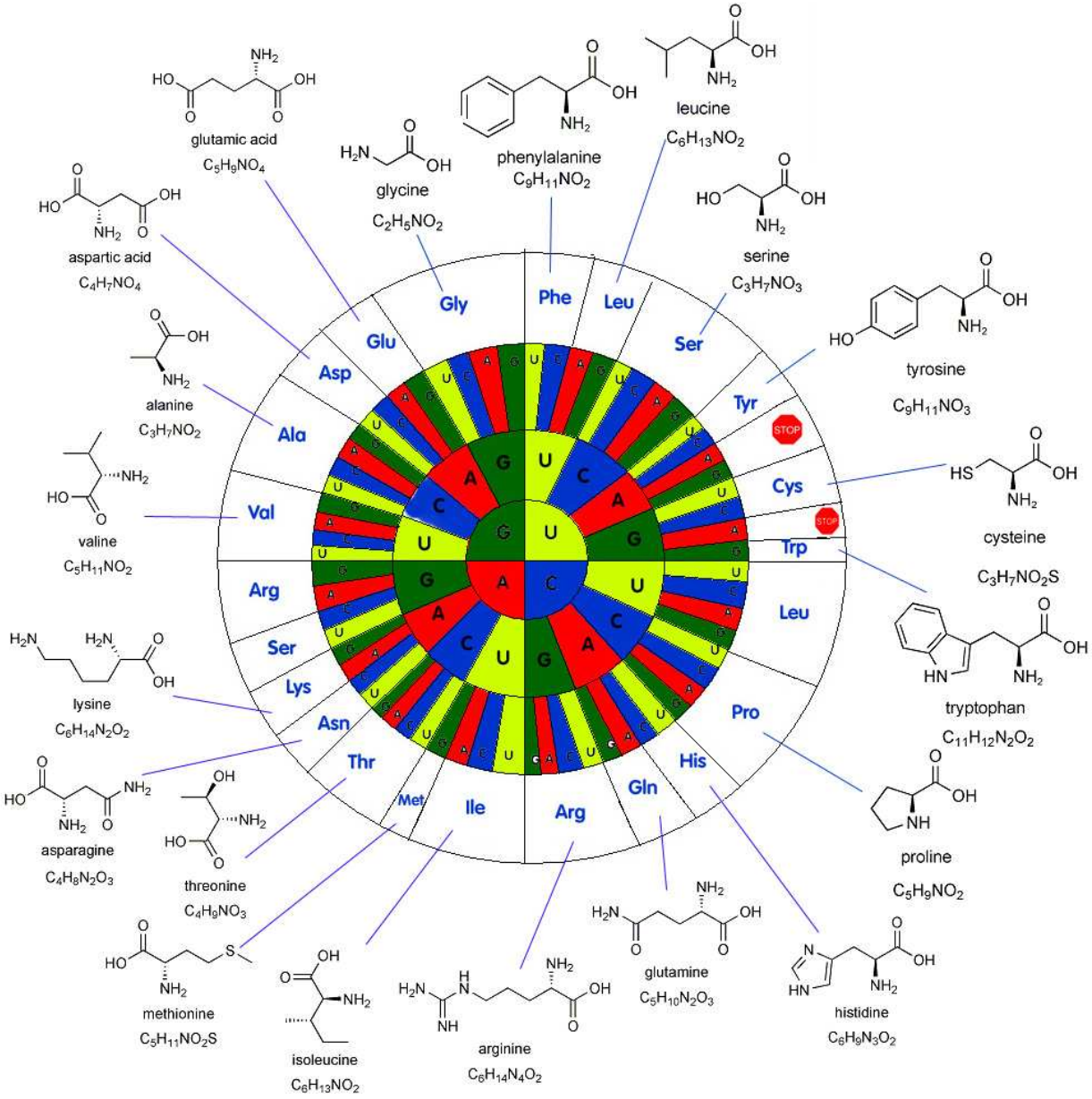


De l'organisme à l'information génétique

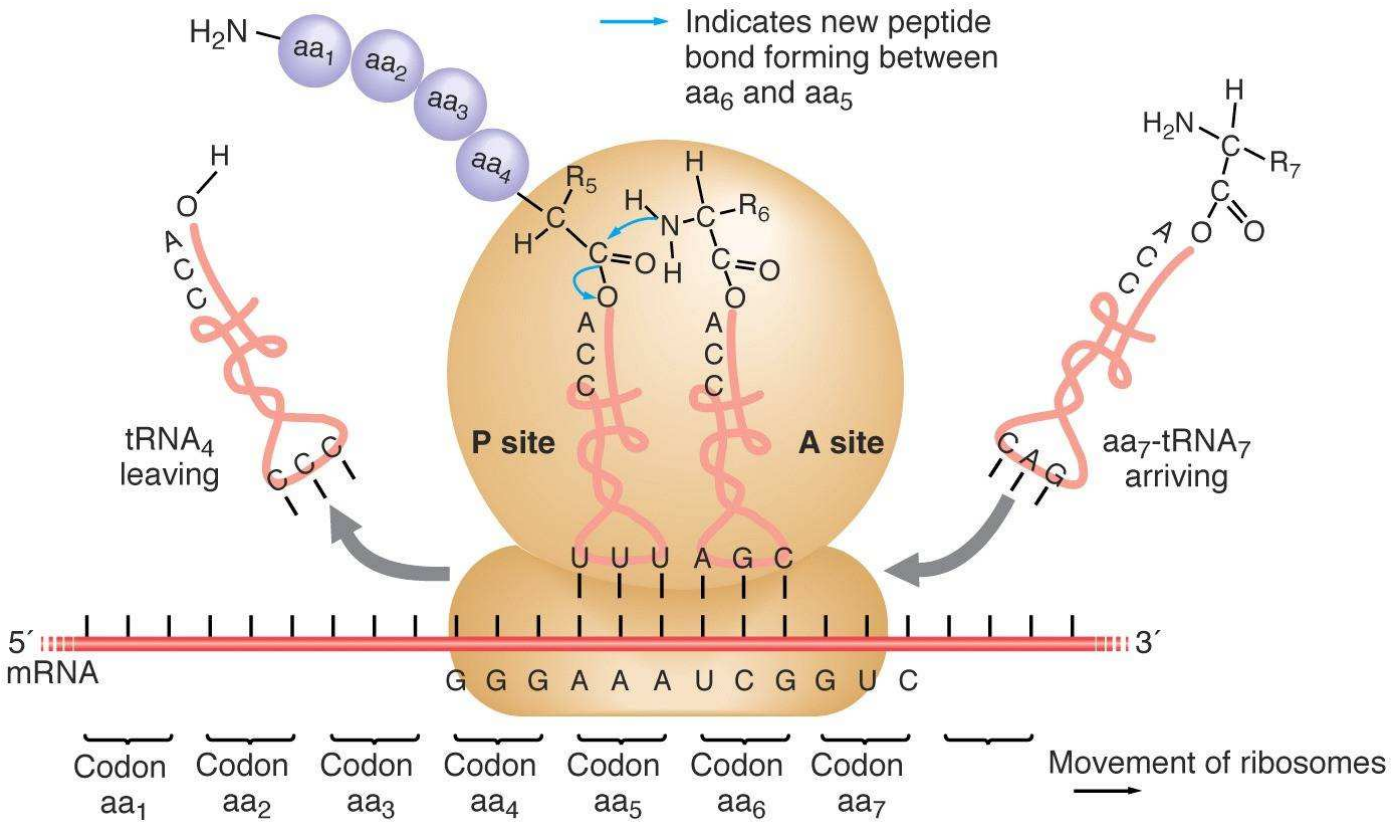




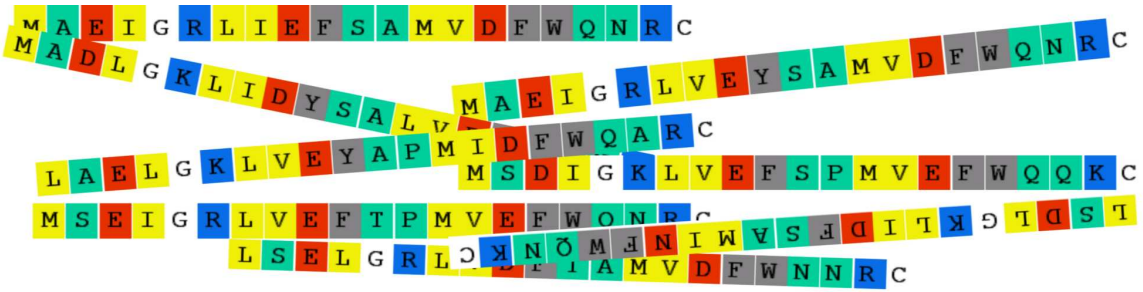
Le code génétique



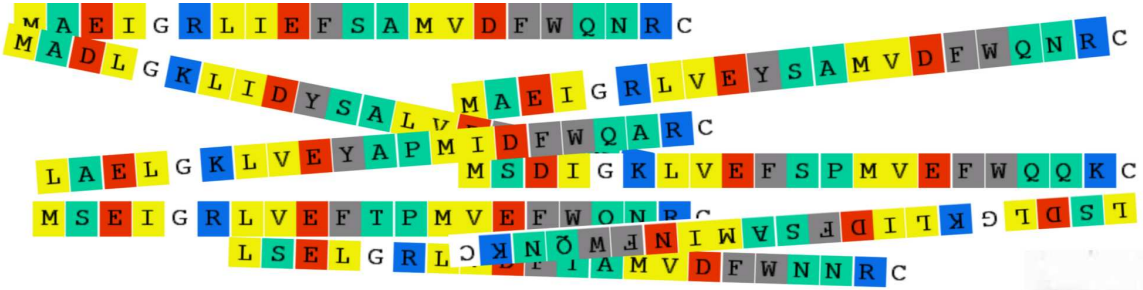
Le ribosome et la traduction des gènes en protéines

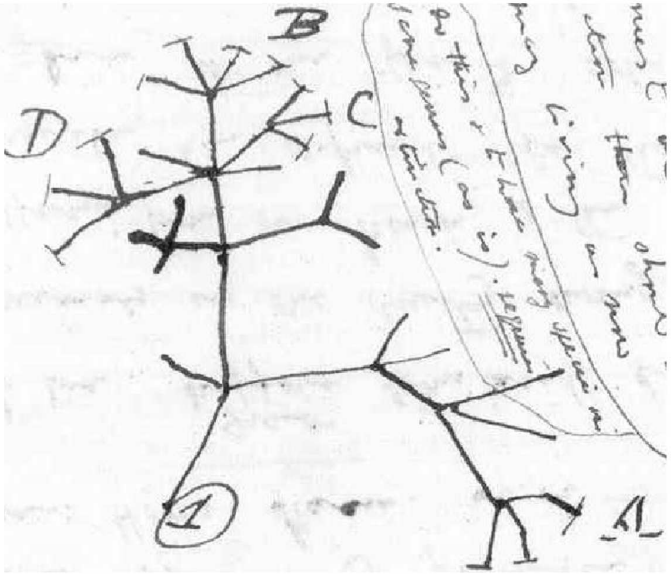


Evolution moléculaire, divergence et homologie

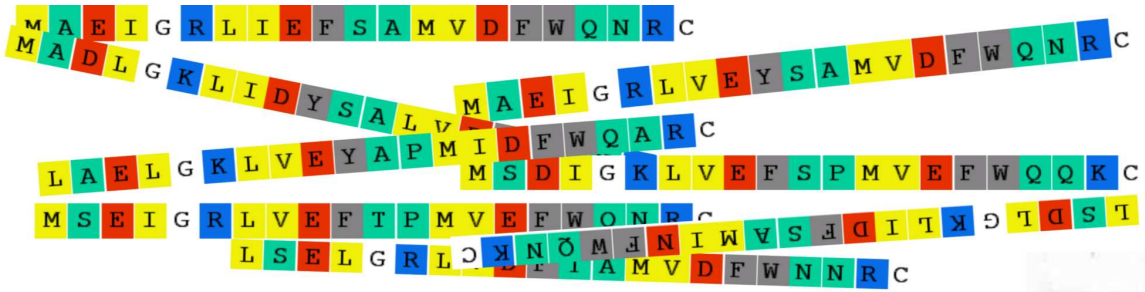


Evolution moléculaire, divergence et homologie

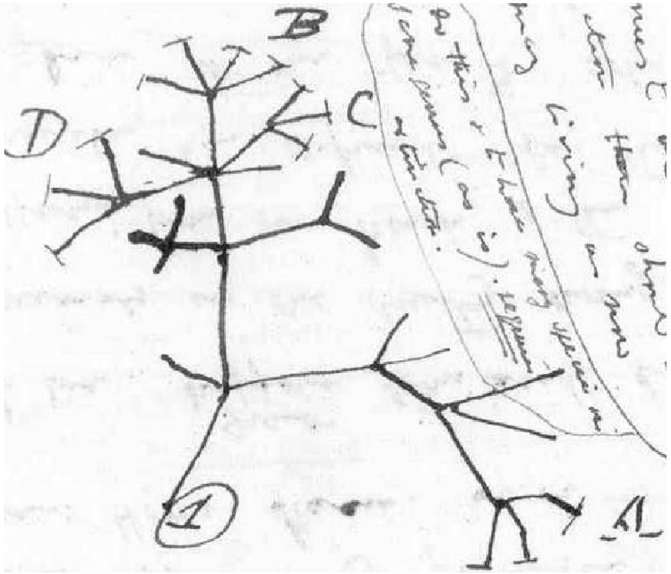




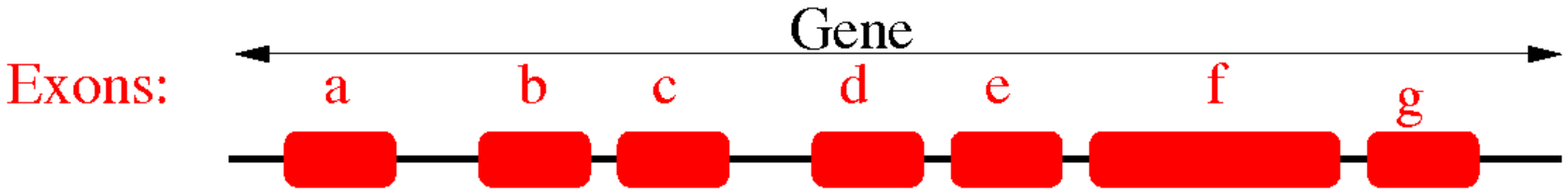
Evolution moléculaire, divergence et homologie



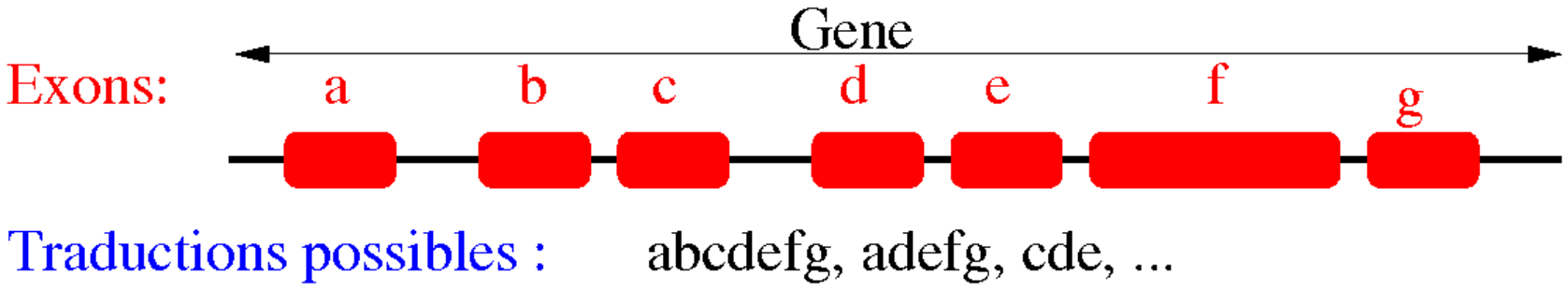
M	A	E	I	G	R	L	I	E	F	S	A	M	V	D	F	W	Q	N	R	C
M	A	E	I	G	R	L	V	E	Y	S	A	M	V	D	F	W	Q	N	R	C
M	A	D	L	G	K	L	I	D	Y	S	A	L	V	D	F	W	Q	N	R	C
M	S	D	I	G	K	L	V	E	F	S	P	M	V	E	F	W	Q	Q	K	C
M	S	E	I	G	R	L	V	E	F	T	P	M	V	E	F	W	Q	N	R	C
L	S	E	L	G	R	L	V	D	F	T	A	M	V	D	F	W	N	N	R	C
L	A	E	L	G	K	L	V	E	Y	A	P	M	I	D	F	W	Q	A	R	C
L	S	D	L	G	K	L	I	D	F	S	A	M	I	N	F	W	Q	N	K	C



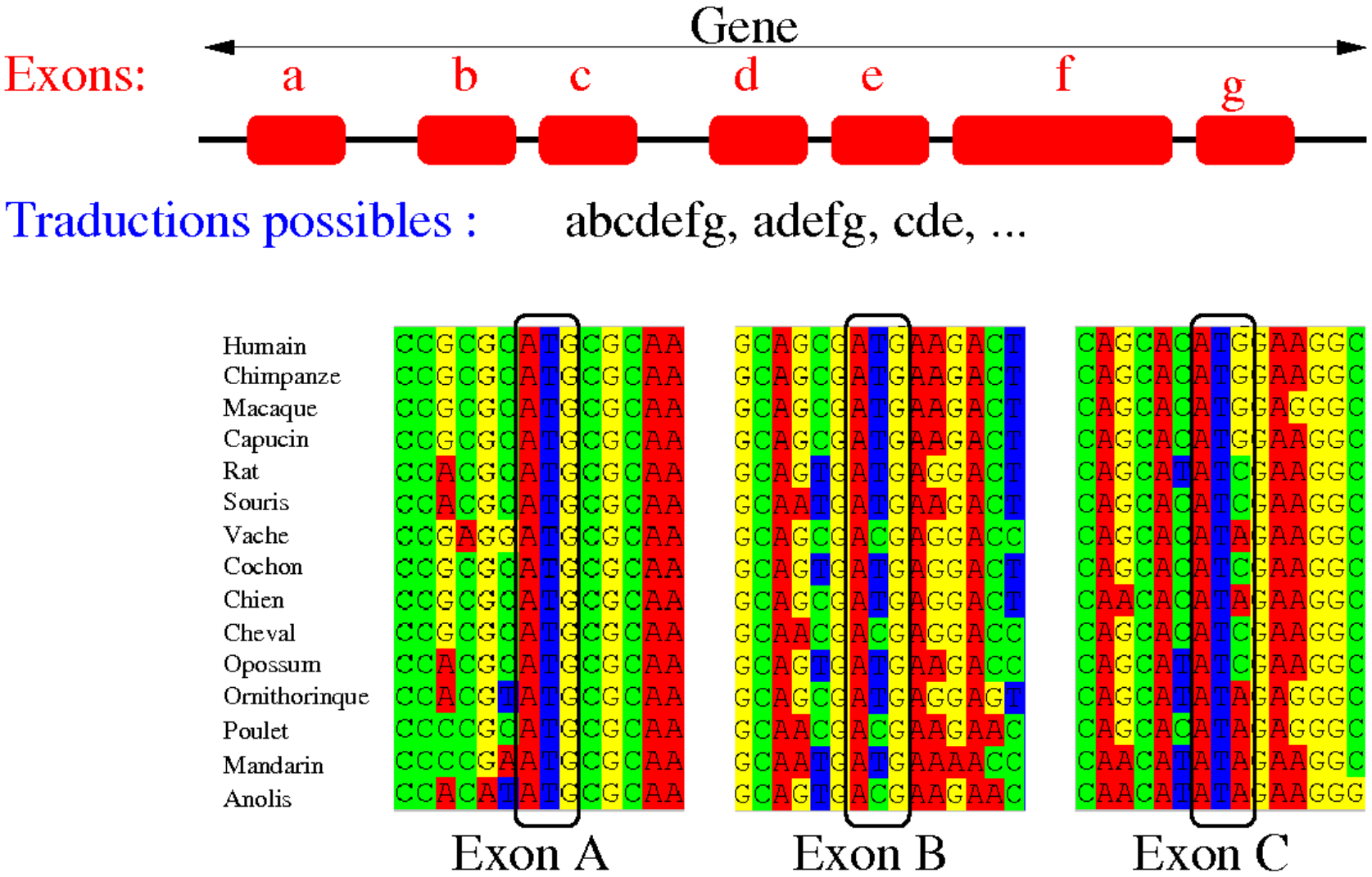
Divergence et homologie, cas des splicing alternatifs



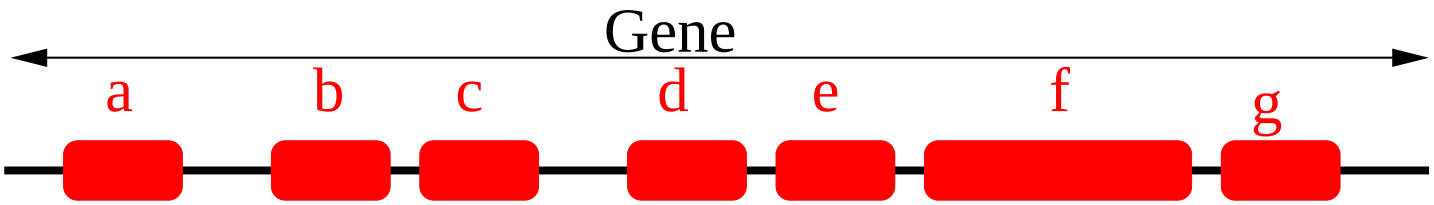
Divergence et homologie, cas des splicing alternatifs



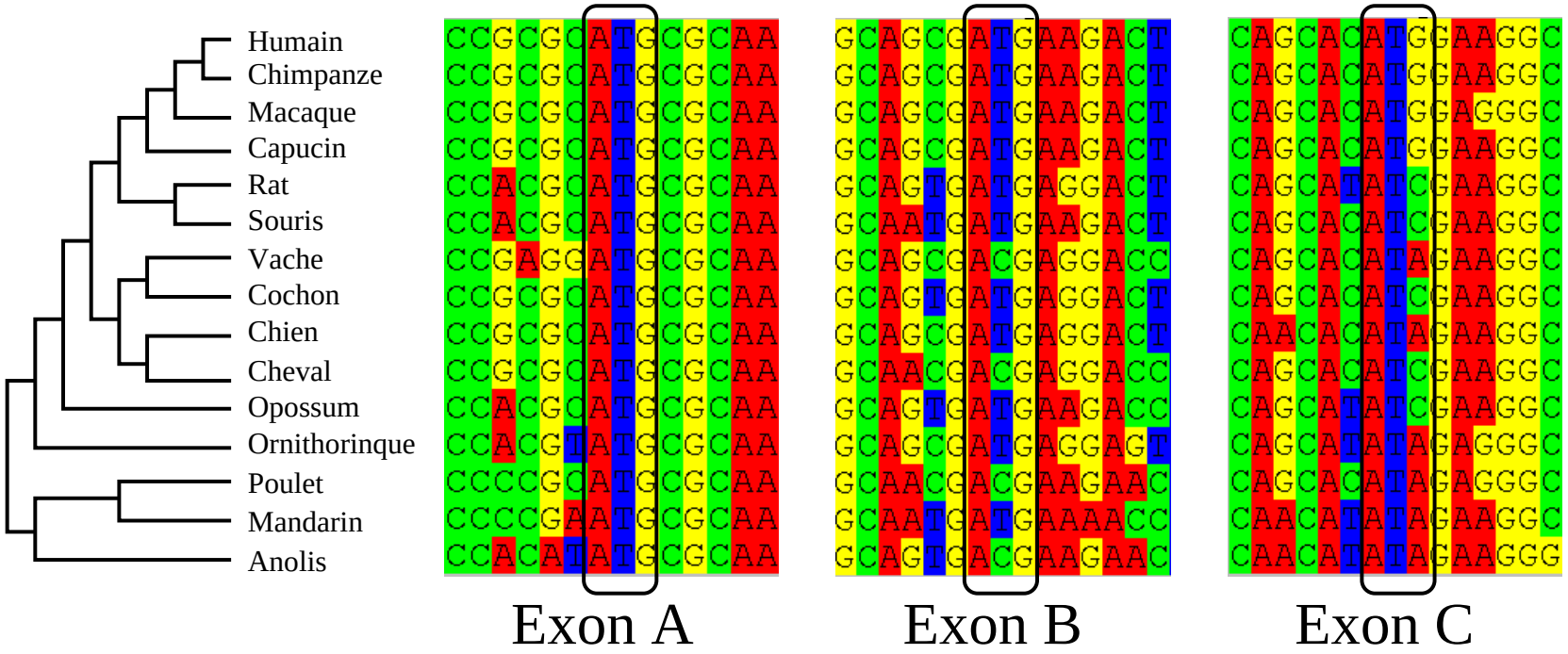
Divergence et homologie, cas des splicing alternatifs



Divergence et homologie, cas des splicing alternatifs

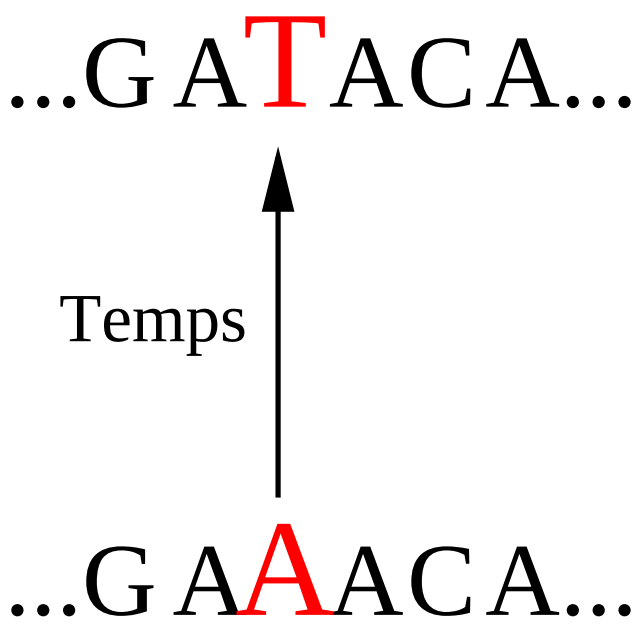


Traductions possibles : abcdefg, adefg, cde, ...



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 $i \rightarrow j$

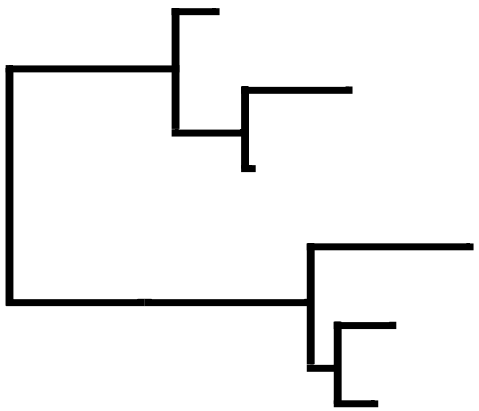
- * définition des cellules diagonales : $Q_{i \rightarrow i} = -\sum_{j \neq i} Q_{i \rightarrow j}$,
- Probabilité d'une substitution en un temps t :

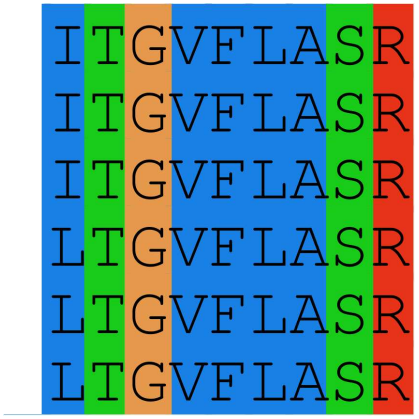
$$P(i \rightarrow j | t) = [e^{t \times Q}]_{i,j}$$

Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$





Données D

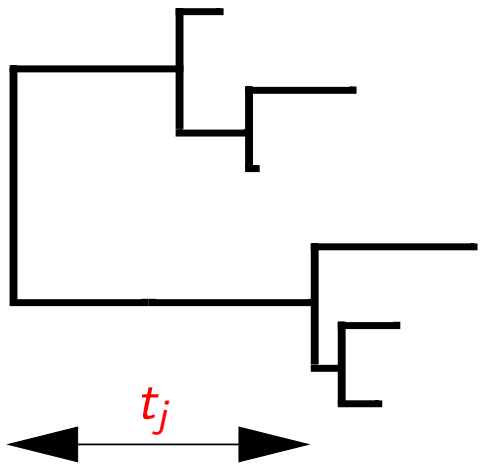
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ

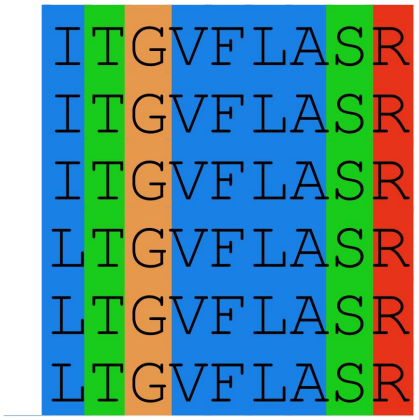
Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$

$\mathbf{t},$





Données D

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

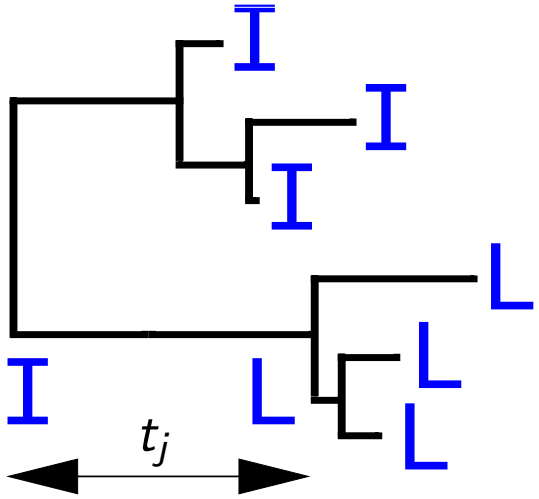
Le modèle phylogénétique standard

Modèle θ

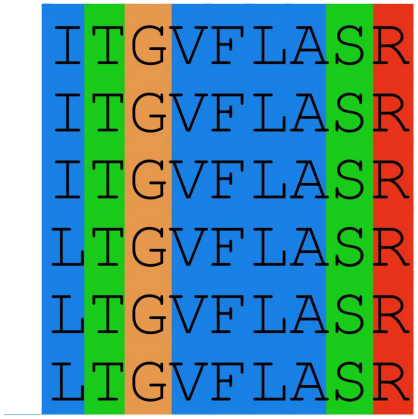
$\theta = \{\tau,$

$\mathbf{t},$

$Q\ \}$



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



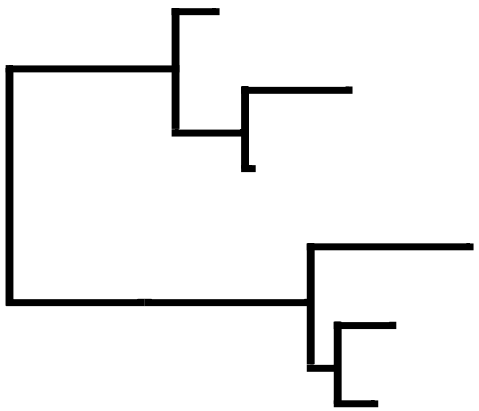
Données D

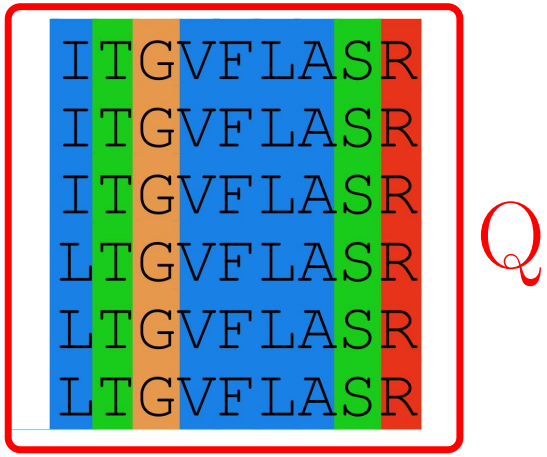
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ, Vitesses d'évolution $\mathbf{t}$,
- Q, générateur Markovien du processus de substitution.

Le modèle phylogénétique standard

Modèle θ

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

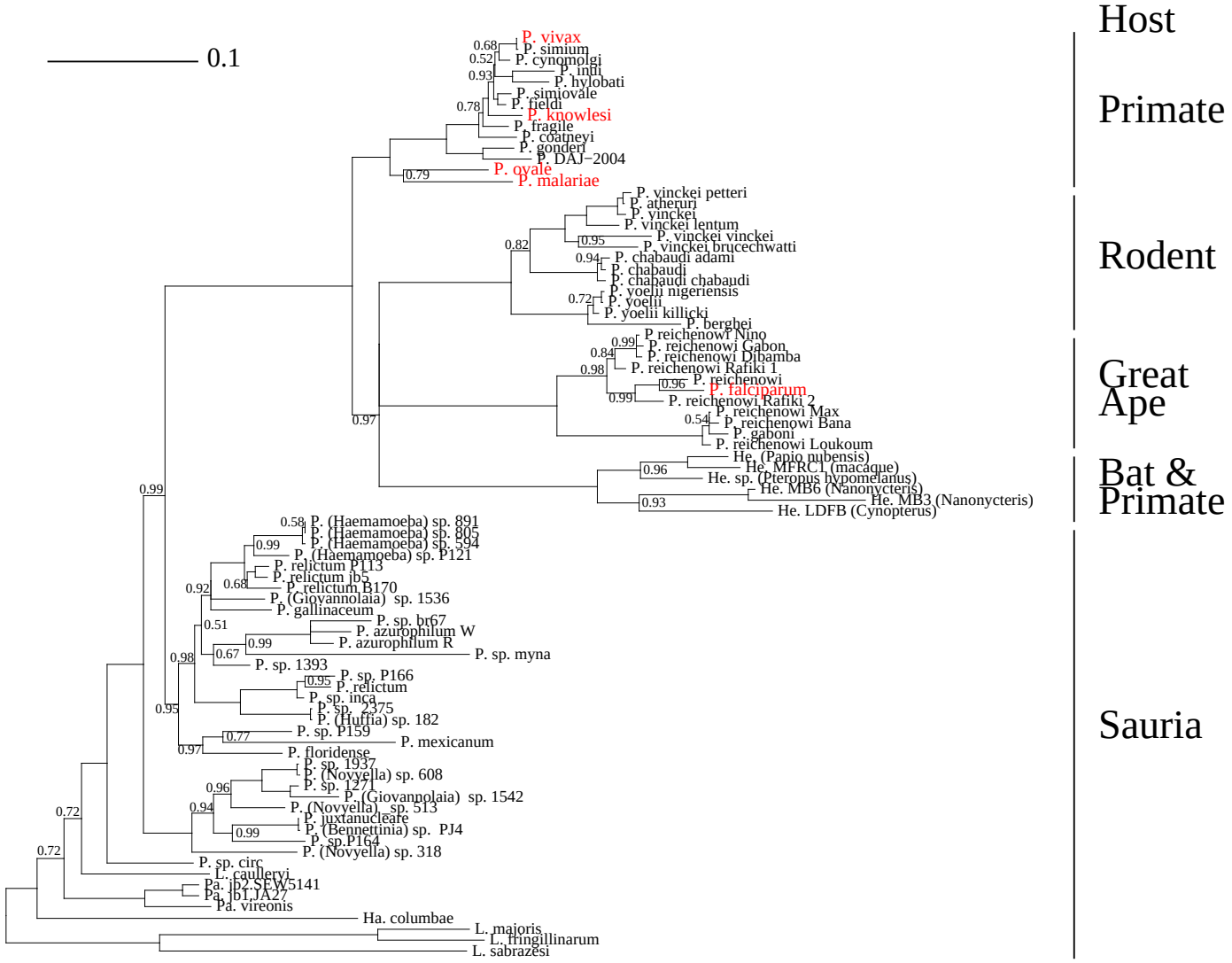




Données D

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

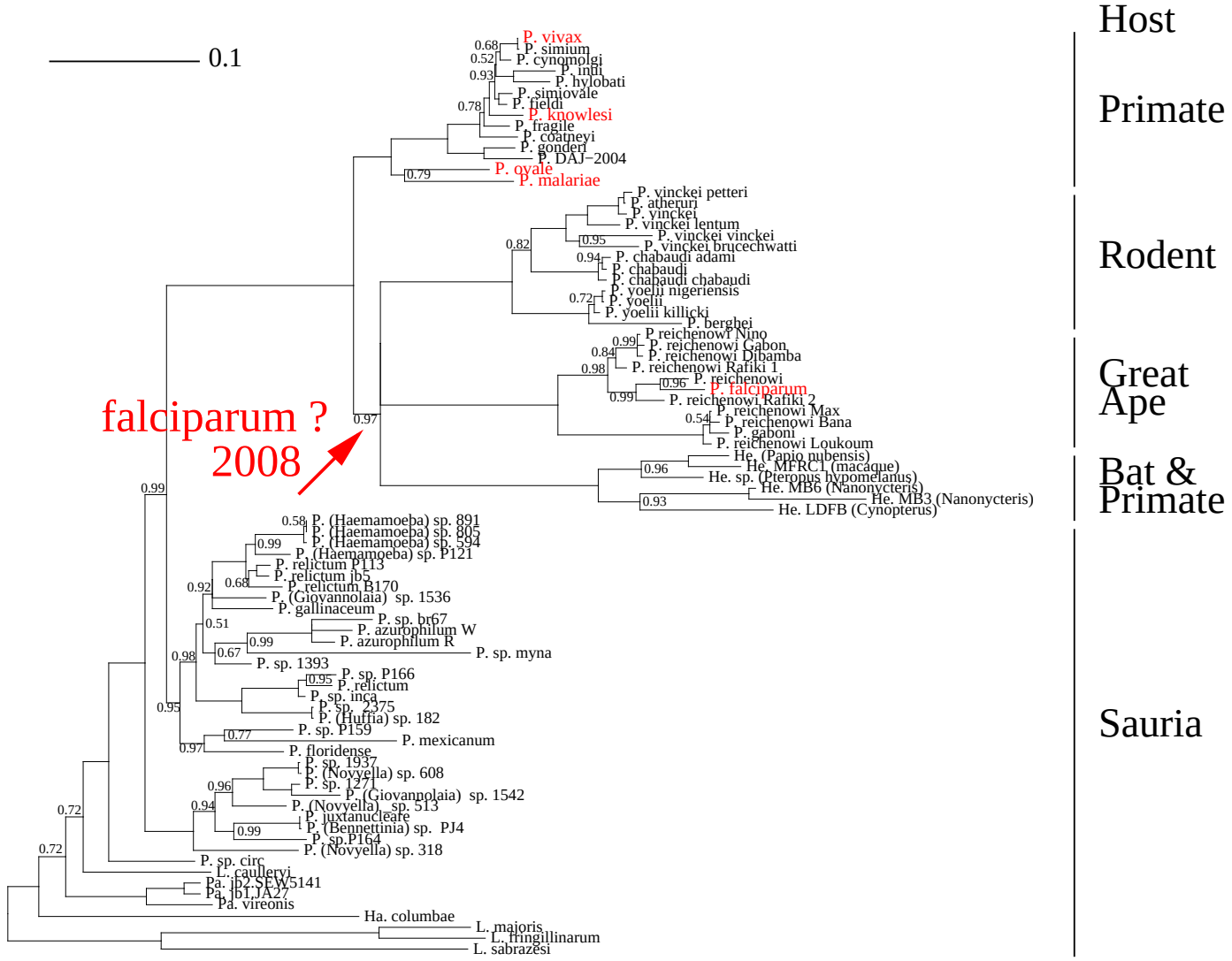
Incertitudes sur la phylogénie des malarias et paléo-écologie



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).

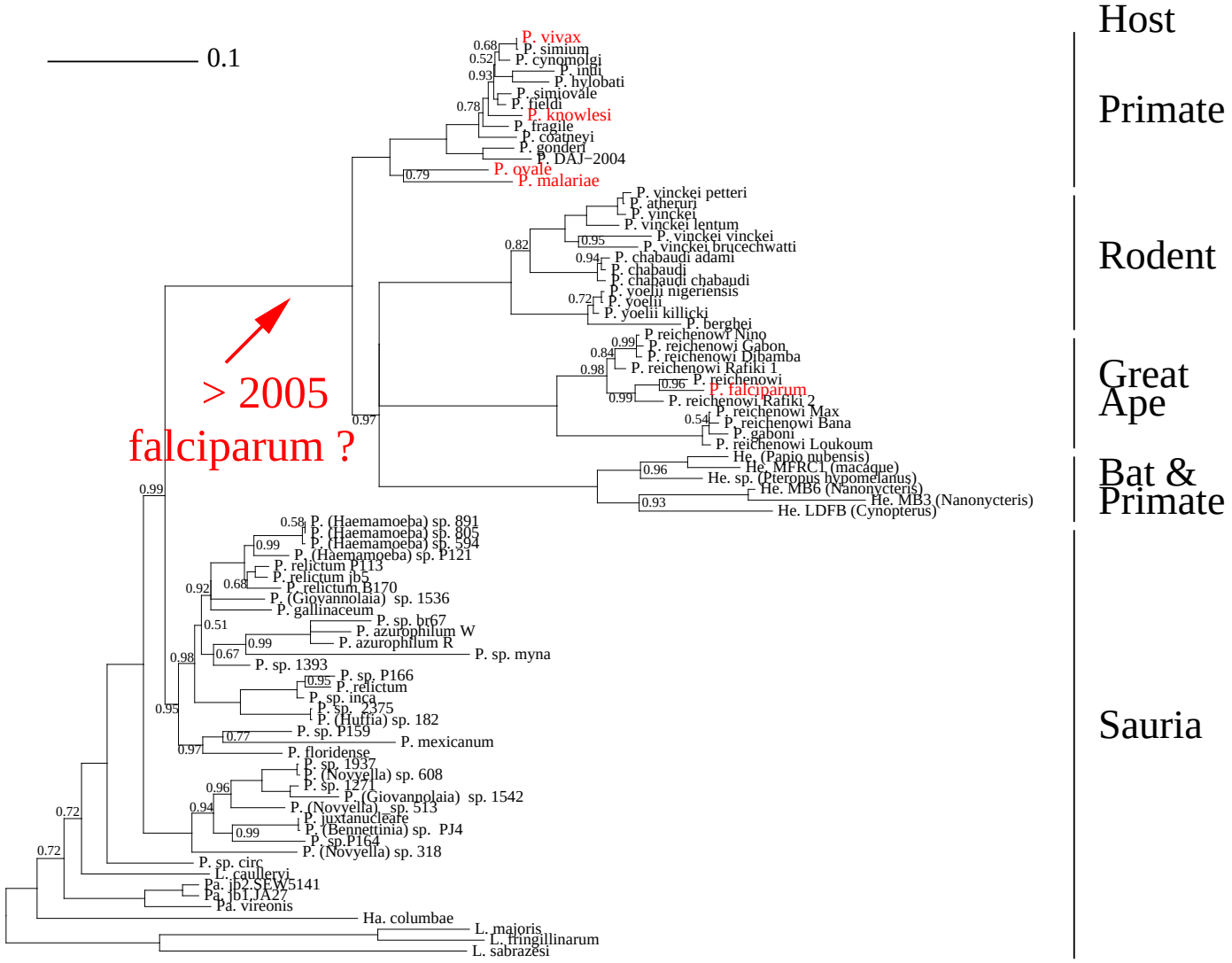
Incertitudes sur la phylogénie des malarias et paléo-écologie



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).

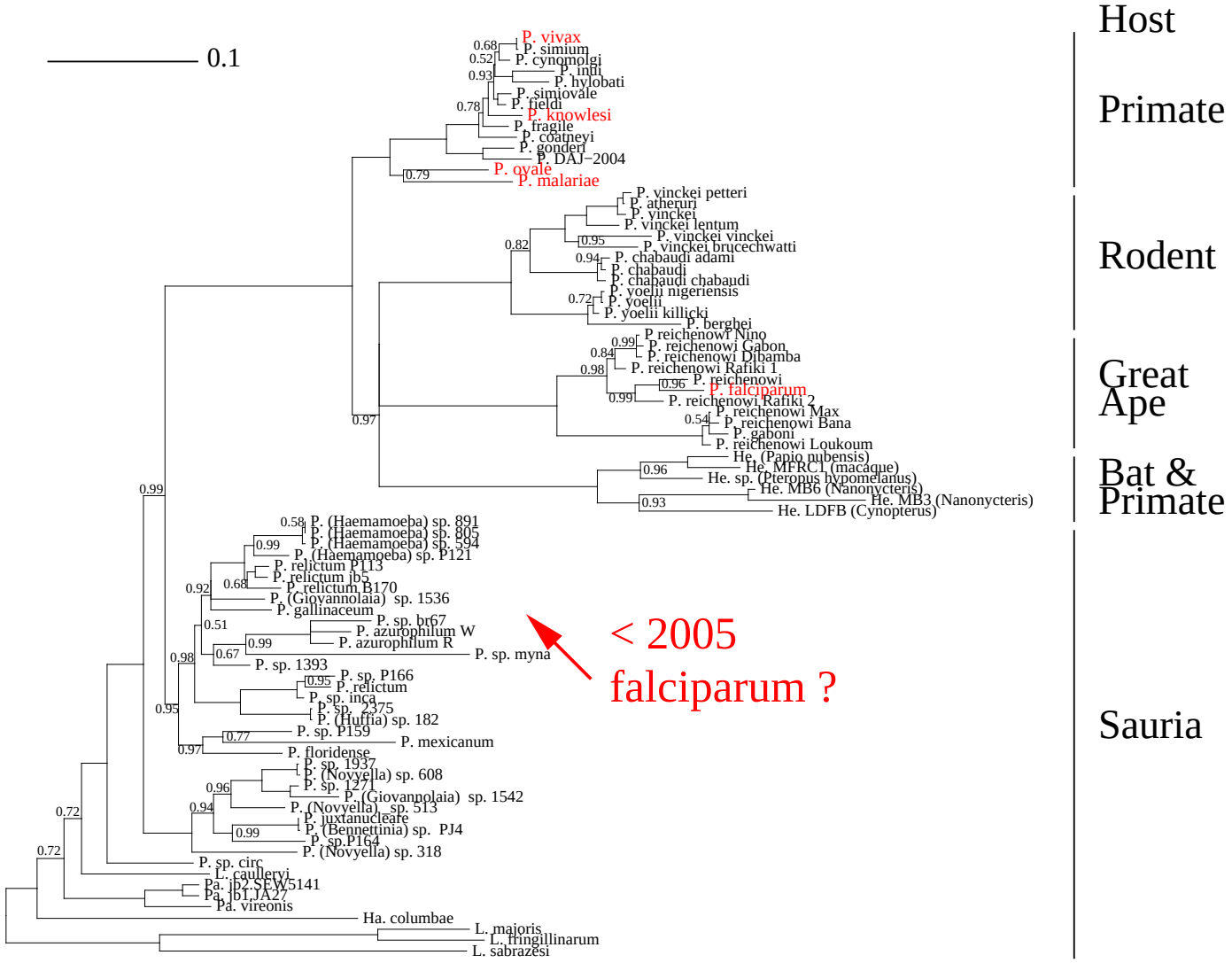
Incertitudes sur la phylogénie des malarias et paléo-écologie



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.

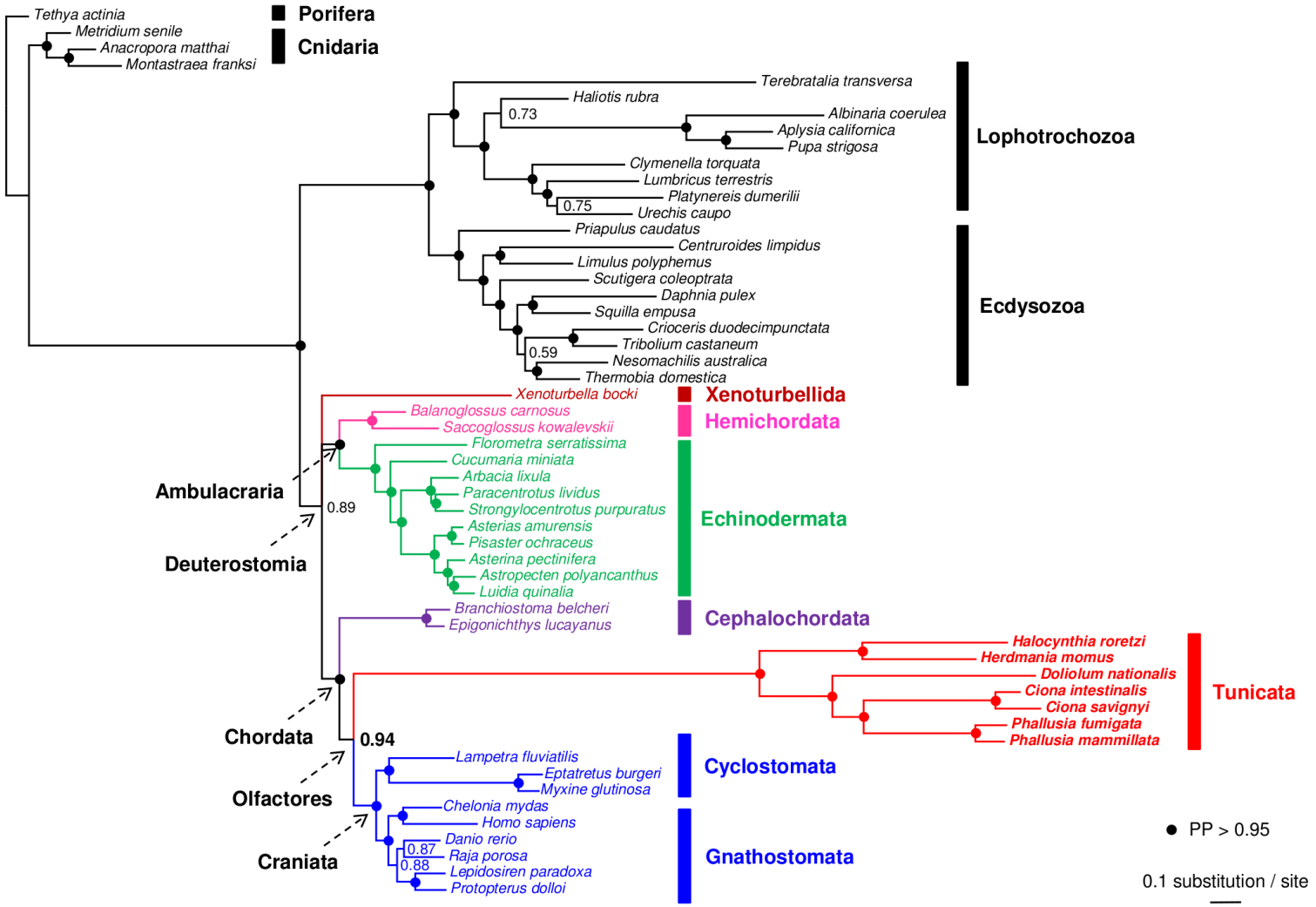
Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

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.

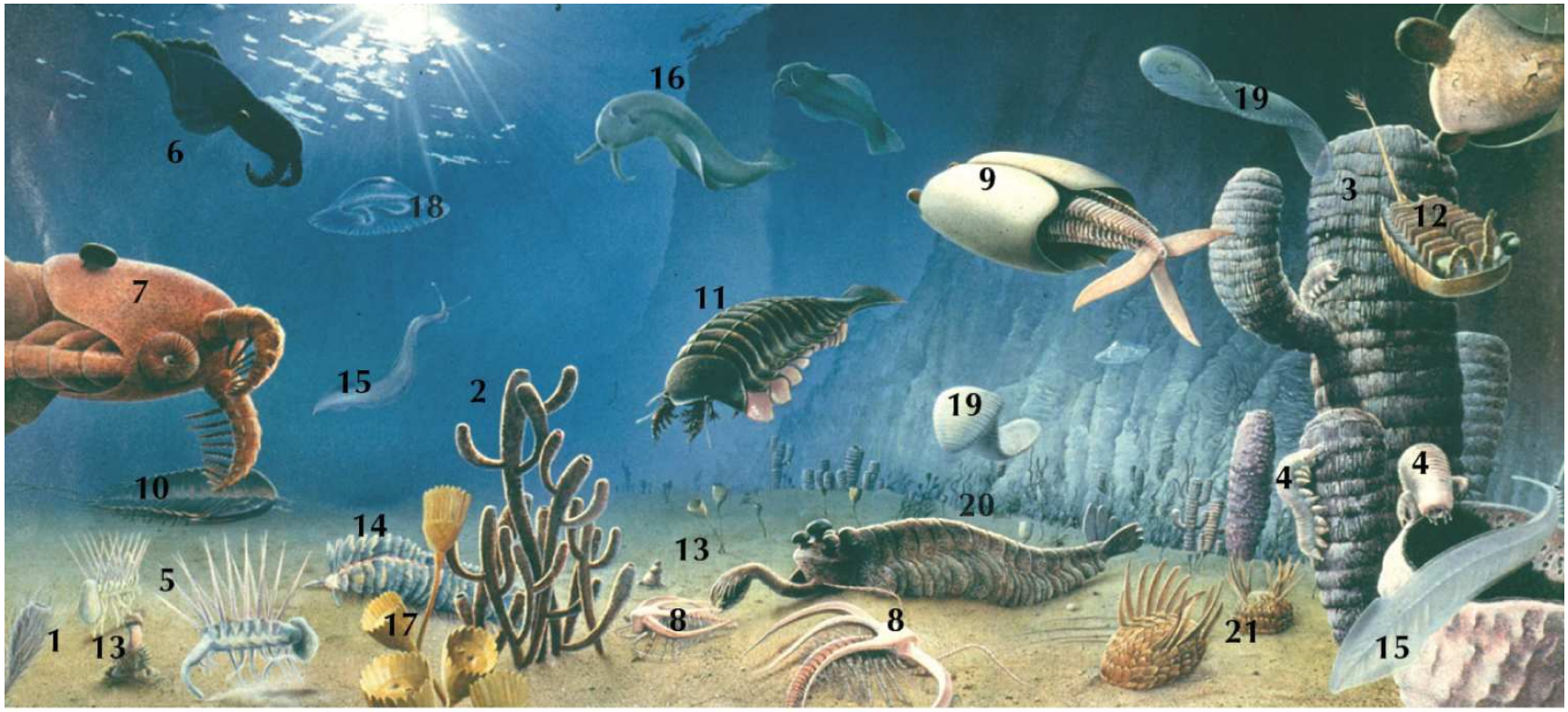
Blanquart & Gascuel. BMC Evolutionary Biology (in revision).

Phylogénie des métazoaires



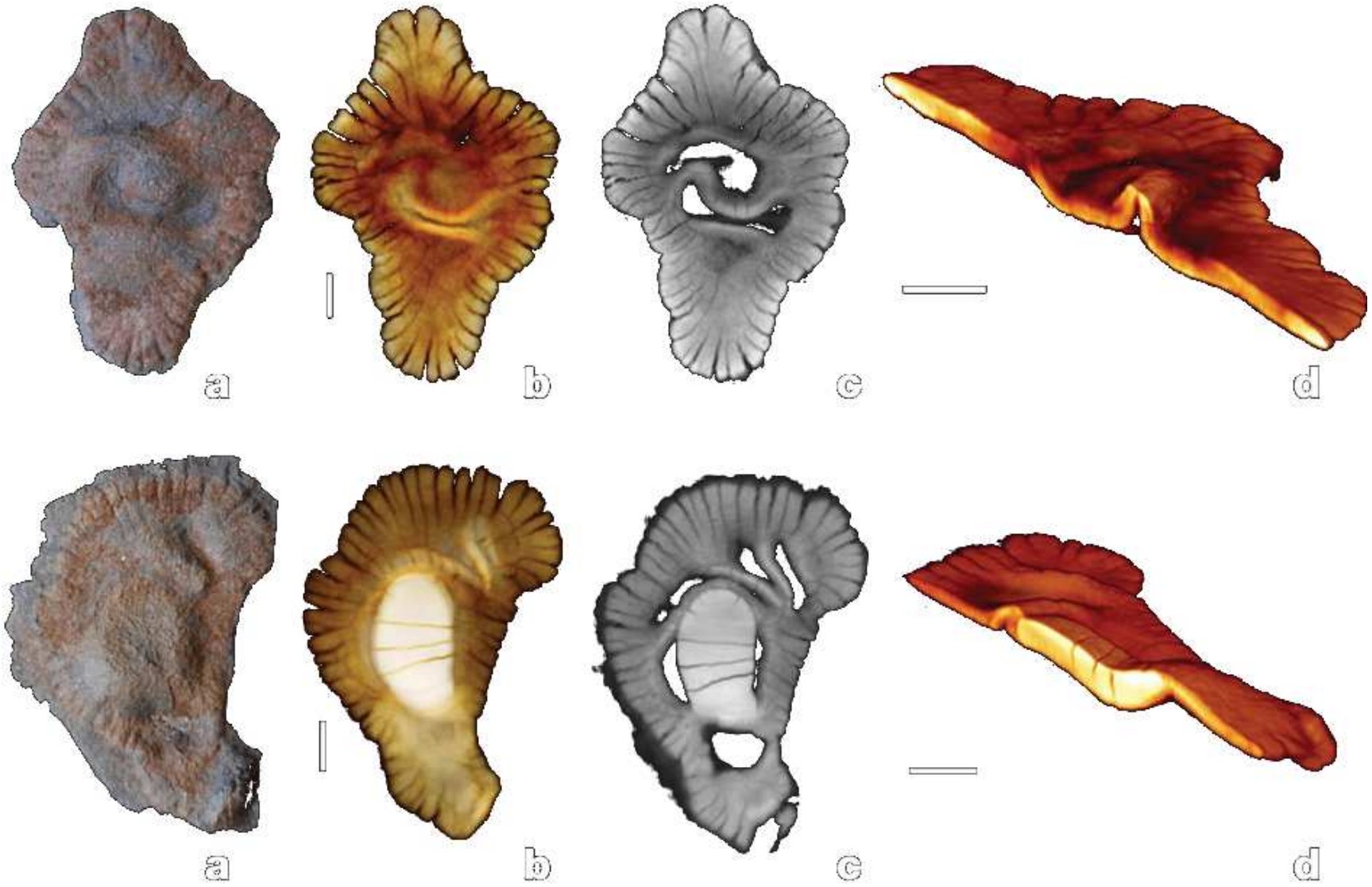
Tunicate mitogenomics and phylogenetics : peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. Singh, Tsagkogeorga, Delsuc, **Blanquart**, Shenkar, Loya, Douzery & Huchon. BMC Evolutionary Biology (2009).

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

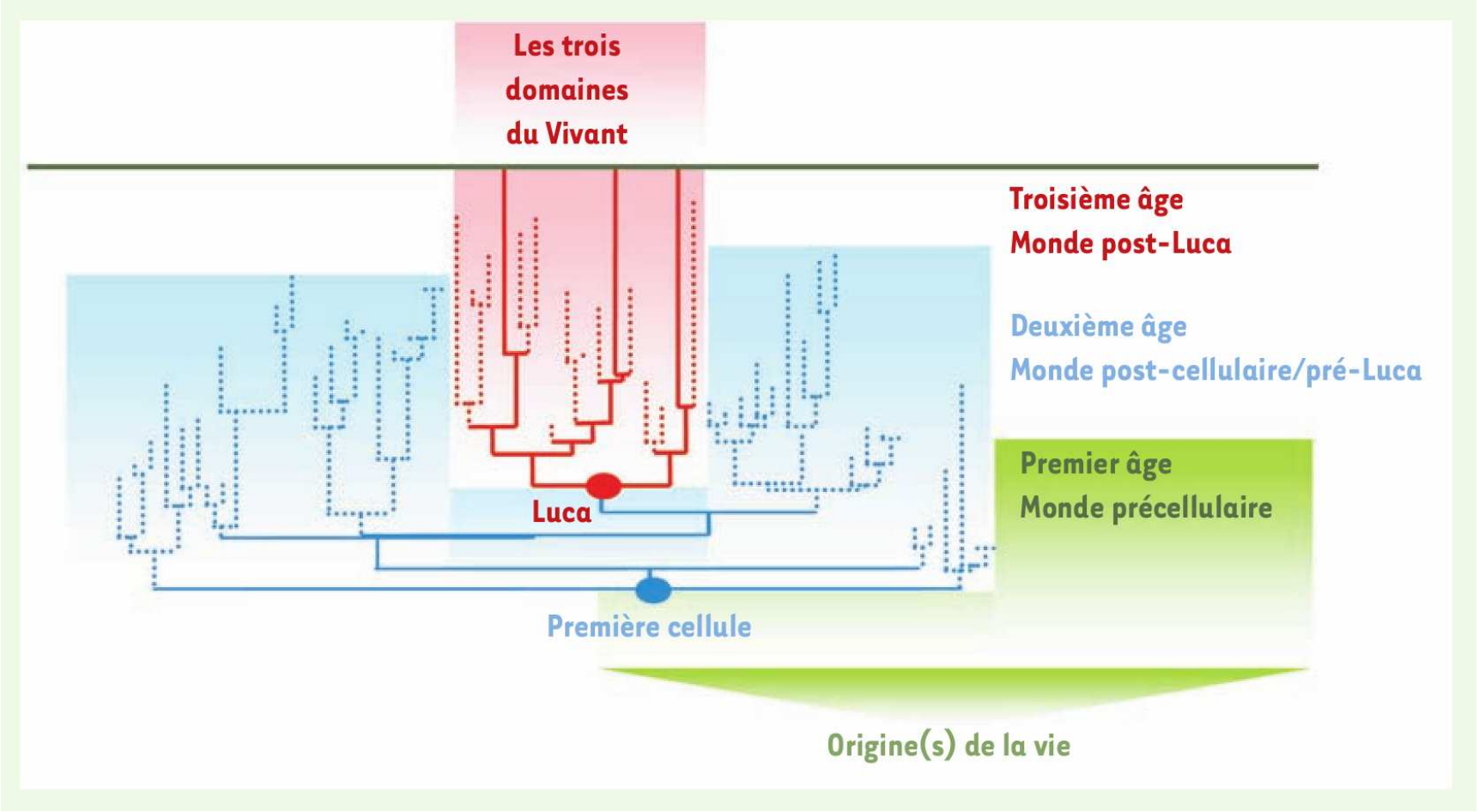


Paléo-faune, Ediacara (Australie) -0.6 milliards d'années

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



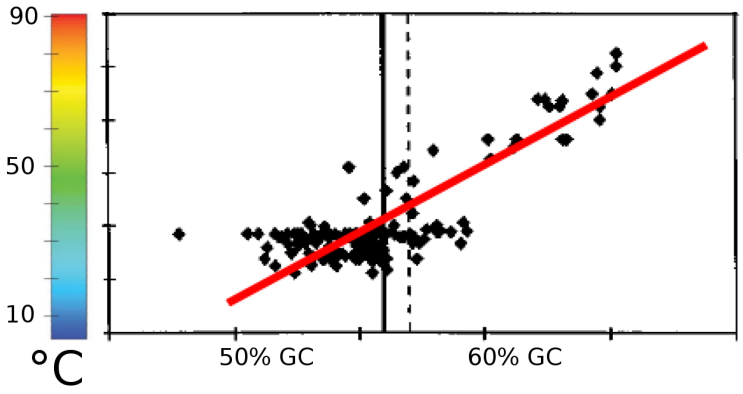
Ecologie des paléo-environnements Archéen et Hadéen



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

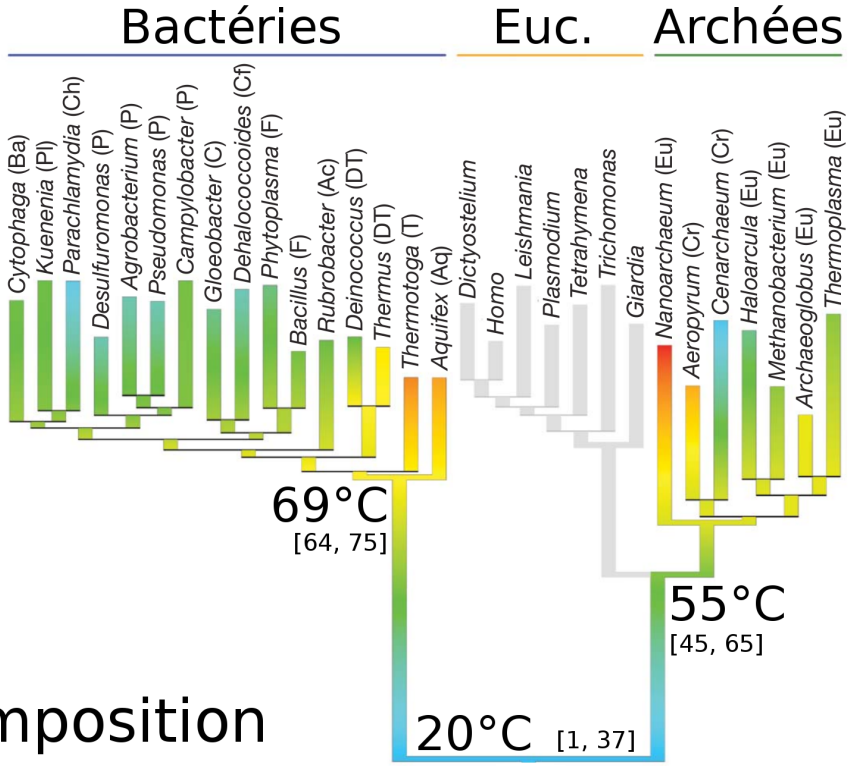
Températures des paléo-environnements

ARN ribosomiques



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

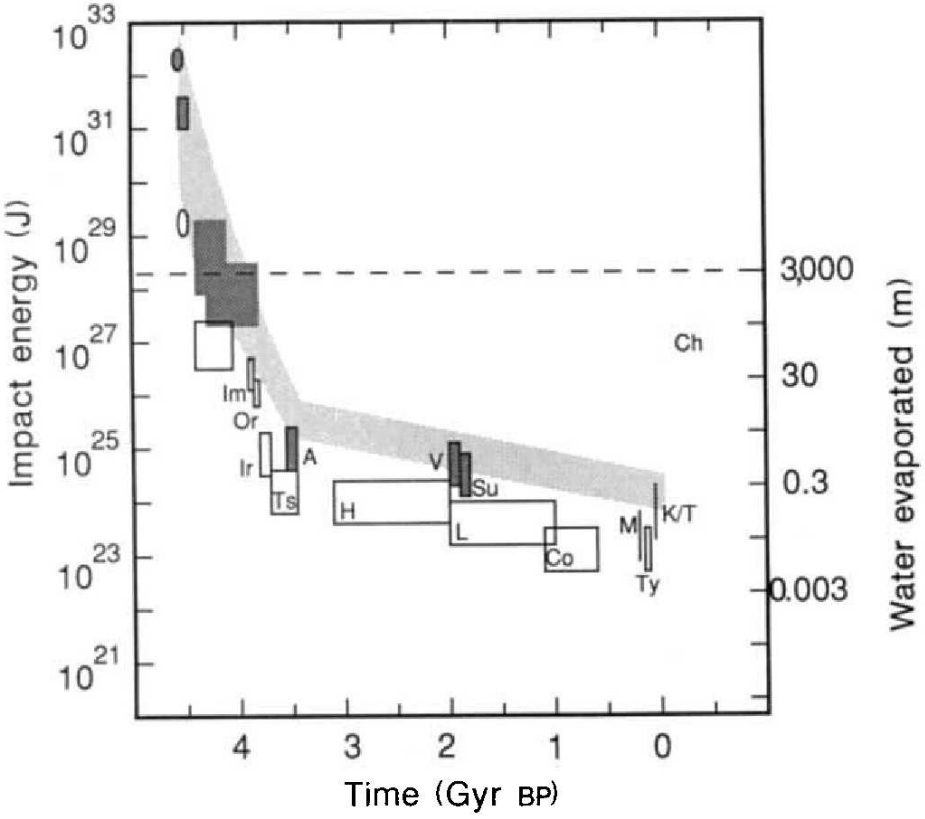
Protéines



Corrélation température/composition
et inférence des compositions ancestrales

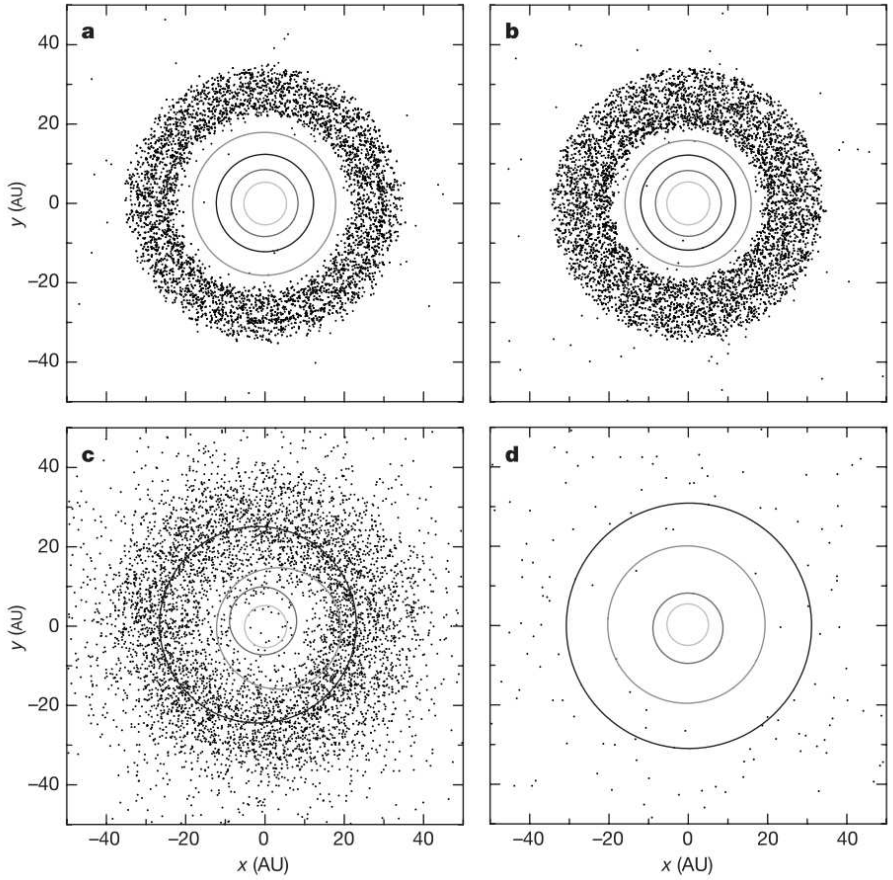
Parallel Adaptations to High Temperatures in the Archean Eon.
Boussau*, **Blanquart***, Necsulea, Lartillot, & Gouy. Nature (2008) (* co premier auteur).

Environnement inter-planétaire



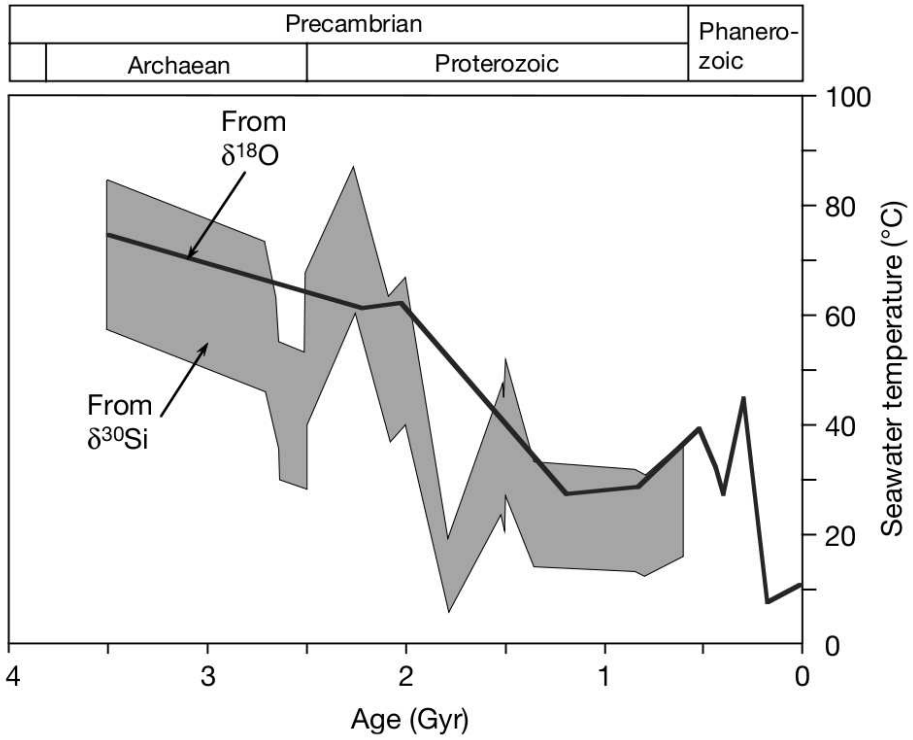
Annihilation of ecosystems by large asteroid impacts on the early Earth.
Sleep, Zahnle, Kasting & Morowitz. Nature (1989).

Le dernier bombardement intense, -3.7 milliards d'années



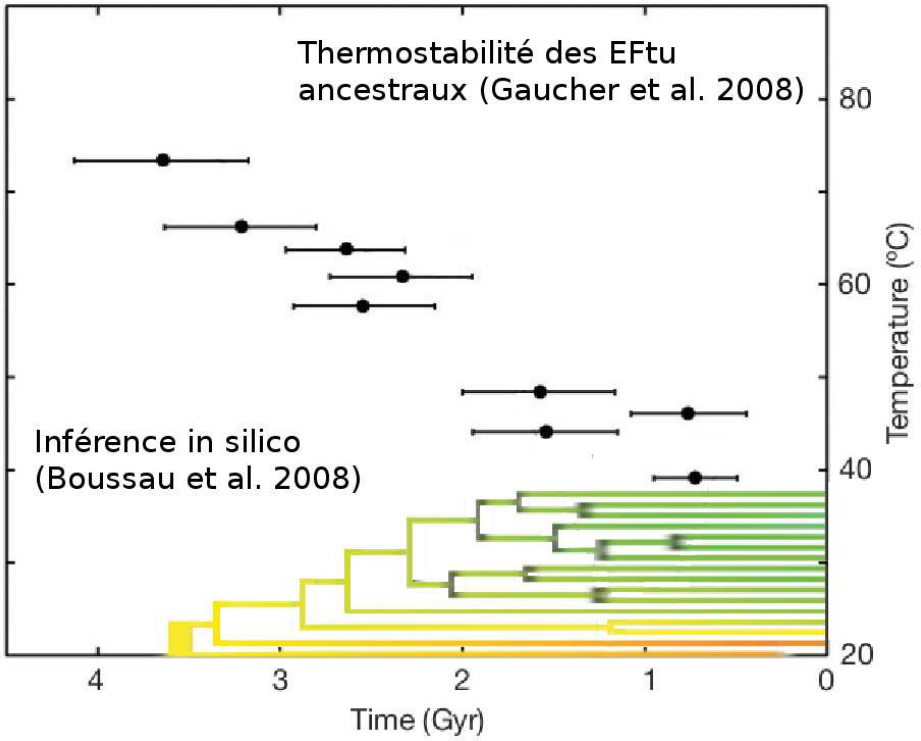
Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. Gomes, Levison, Tsiganis & Morbidelli. Nature (2005).

Température des paléo-océans



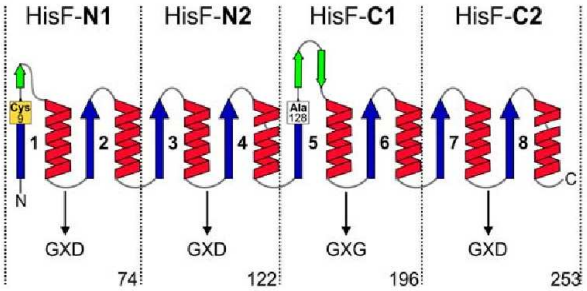
Palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. Robert & Chaussidon. Nature (2006).

Evolution de la thermophilie bactérienne



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

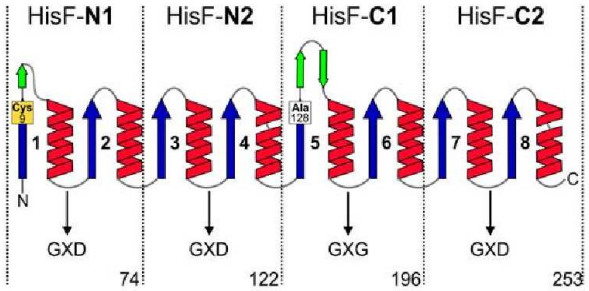
Résurrection de gènes ancestraux



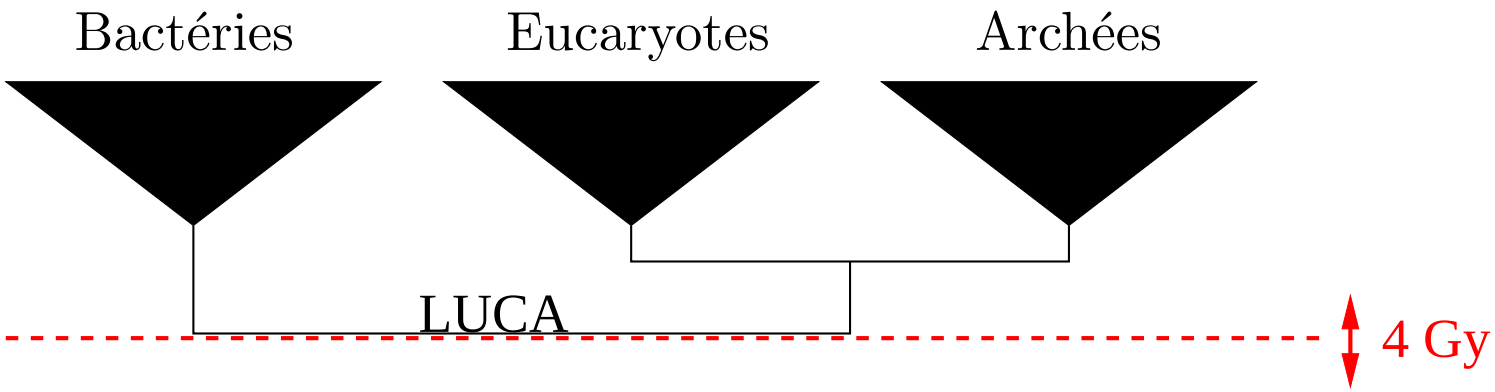
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA

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



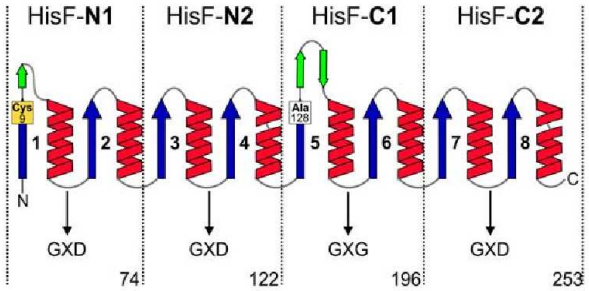
Richter et al. (2010) J Mol. Biol.
Duplications/fusion de dimères $(\beta\alpha)_2$ antérieures à LUCA



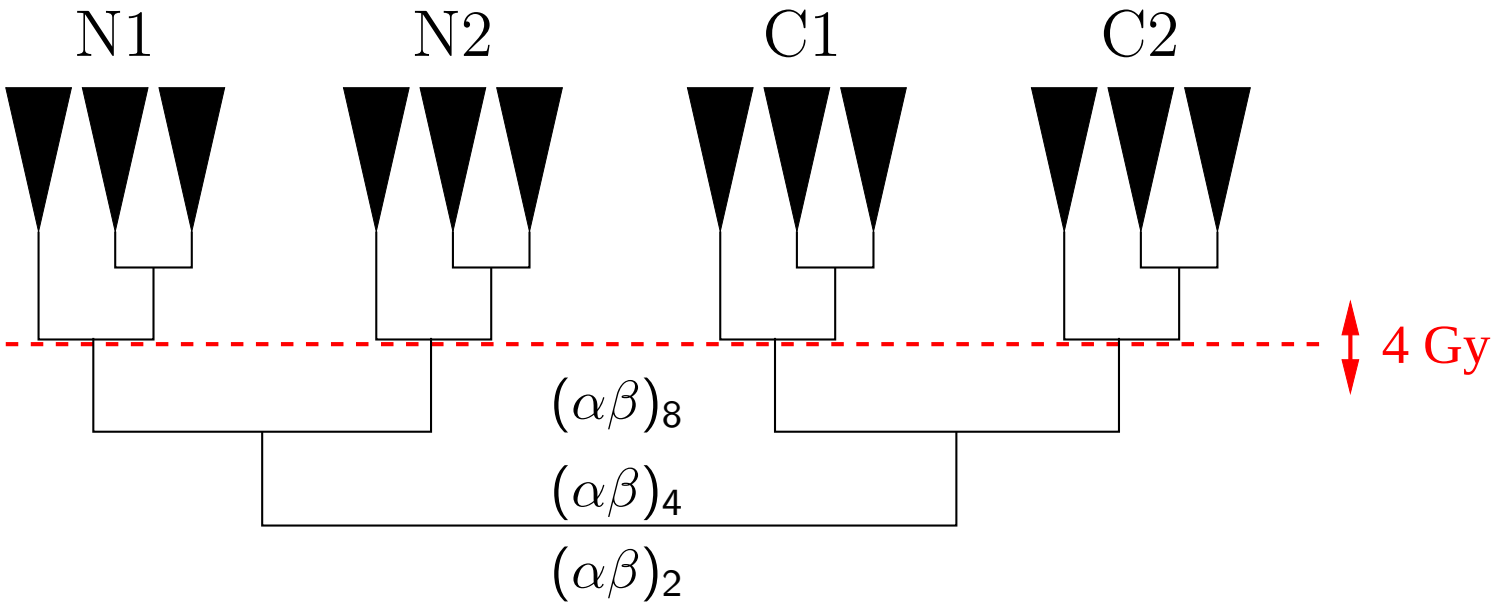
Arbre shématique des 8-mères $(\alpha\beta)_8$

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.

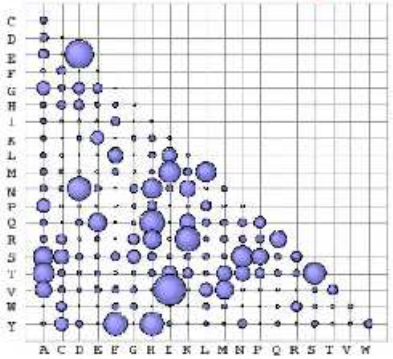
- Question : Quelle est la sensibilité des résurrections moléculaires aux incertitudes statistiques liées aux inférences ?
- Question : Quelle est la fiabilité statistique des inférences extrêmement profondes ?
- Question : Quel est "l'horizon entropique" au delà duquel tout signal phylogénétique apparaît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

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

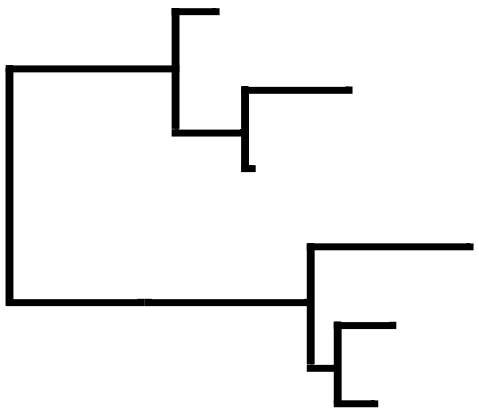


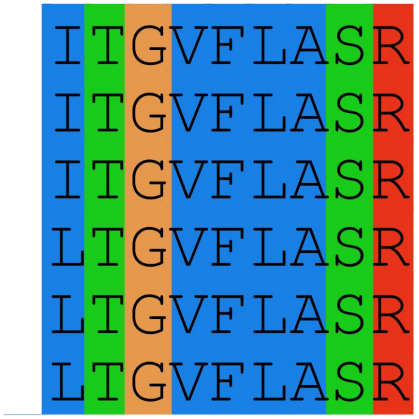
$$\begin{cases} Q_{l \neq m} = \pi_m \rho_{lm} \\ Q_{ll} = -\sum_{m \neq l} Q_{lm} \end{cases}$$

Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$





Données D

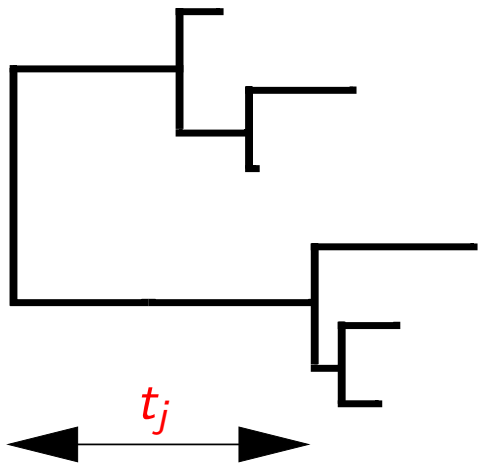
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ

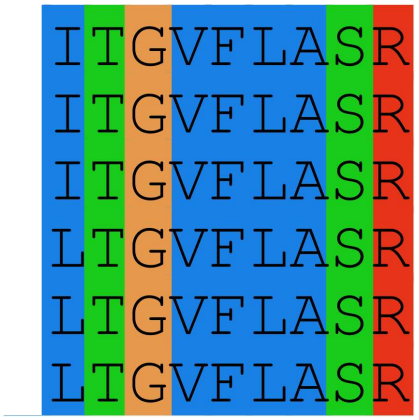
Le modèle phylogénétique standard

Modèle θ

$\theta = \{\tau,$

$\mathbf{t},$





Données D

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

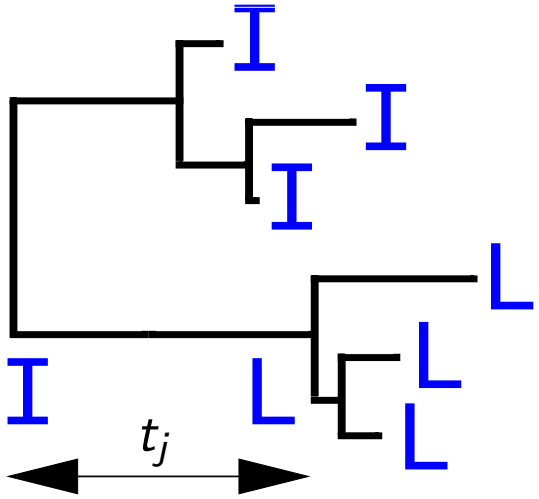
Le modèle phylogénétique standard

Modèle θ

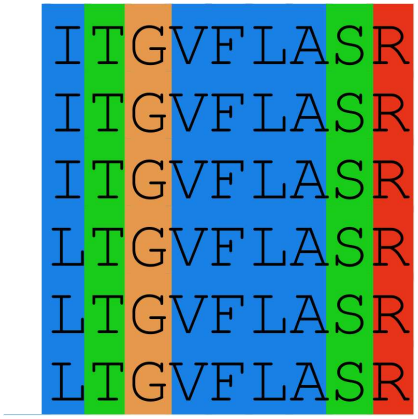
$\theta = \{\tau,$

$\mathbf{t},$

$Q\ \}$



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



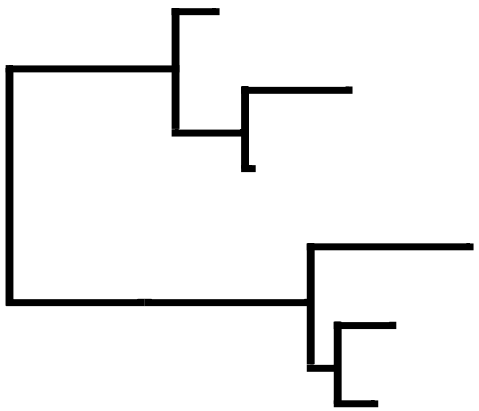
Données D

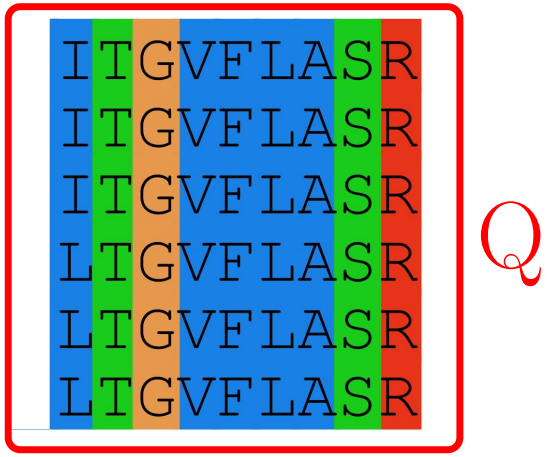
- Données D, ensemble de séquences homologues,
- Modèle : Topologie τ, Vitesses d'évolution $\mathbf{t}$,
- Q, générateur Markovien du processus de substitution.

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.

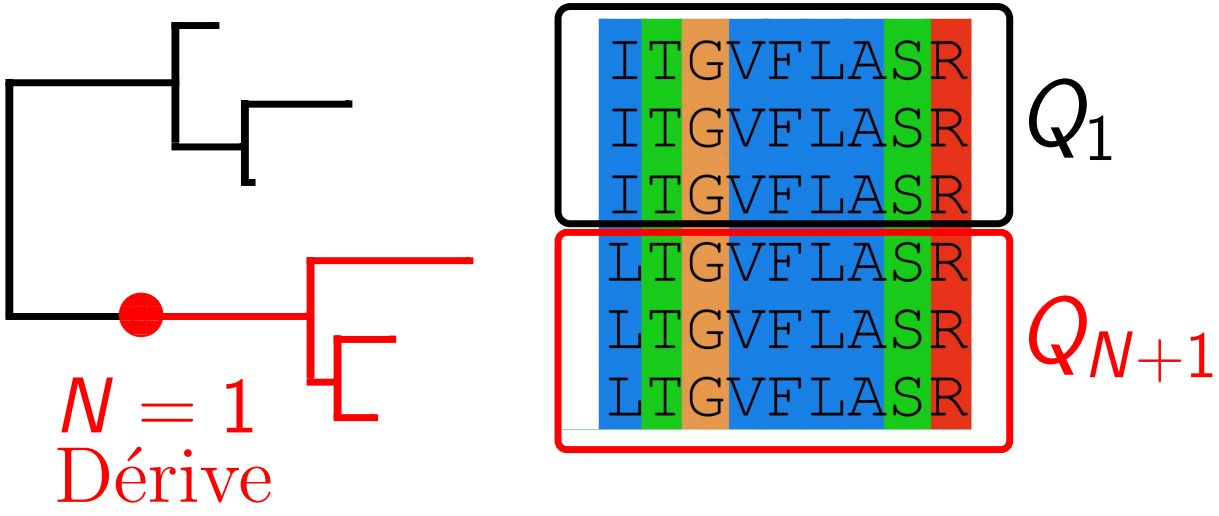
Relaxer l'hypothèse d'homogénéité

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).

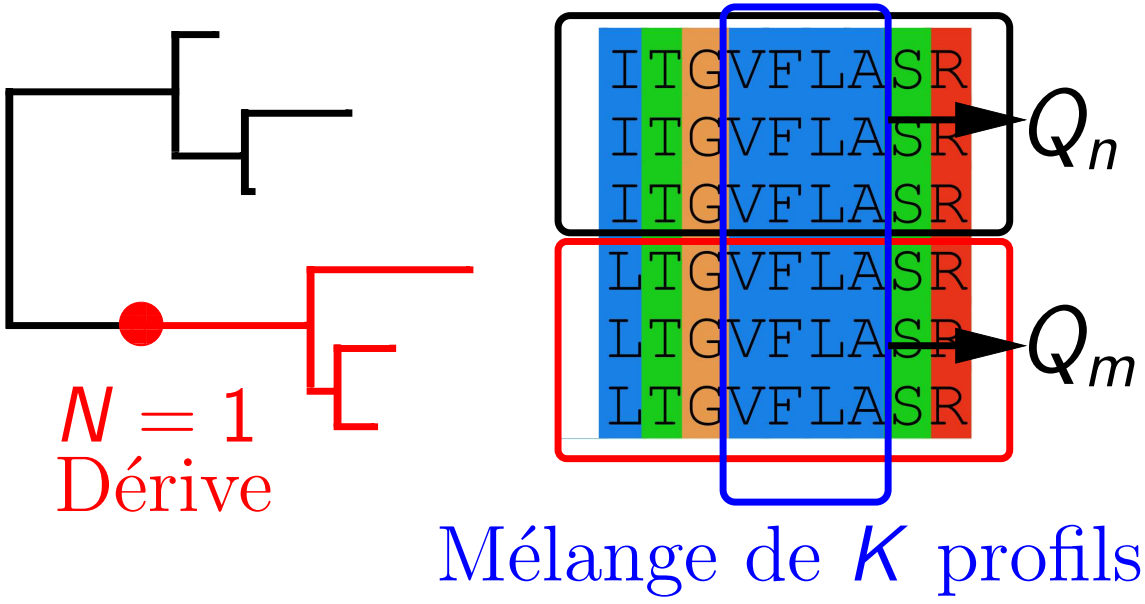
Relaxer l'hypothèse d'homogénéité

Modèle θ

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

N, K

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



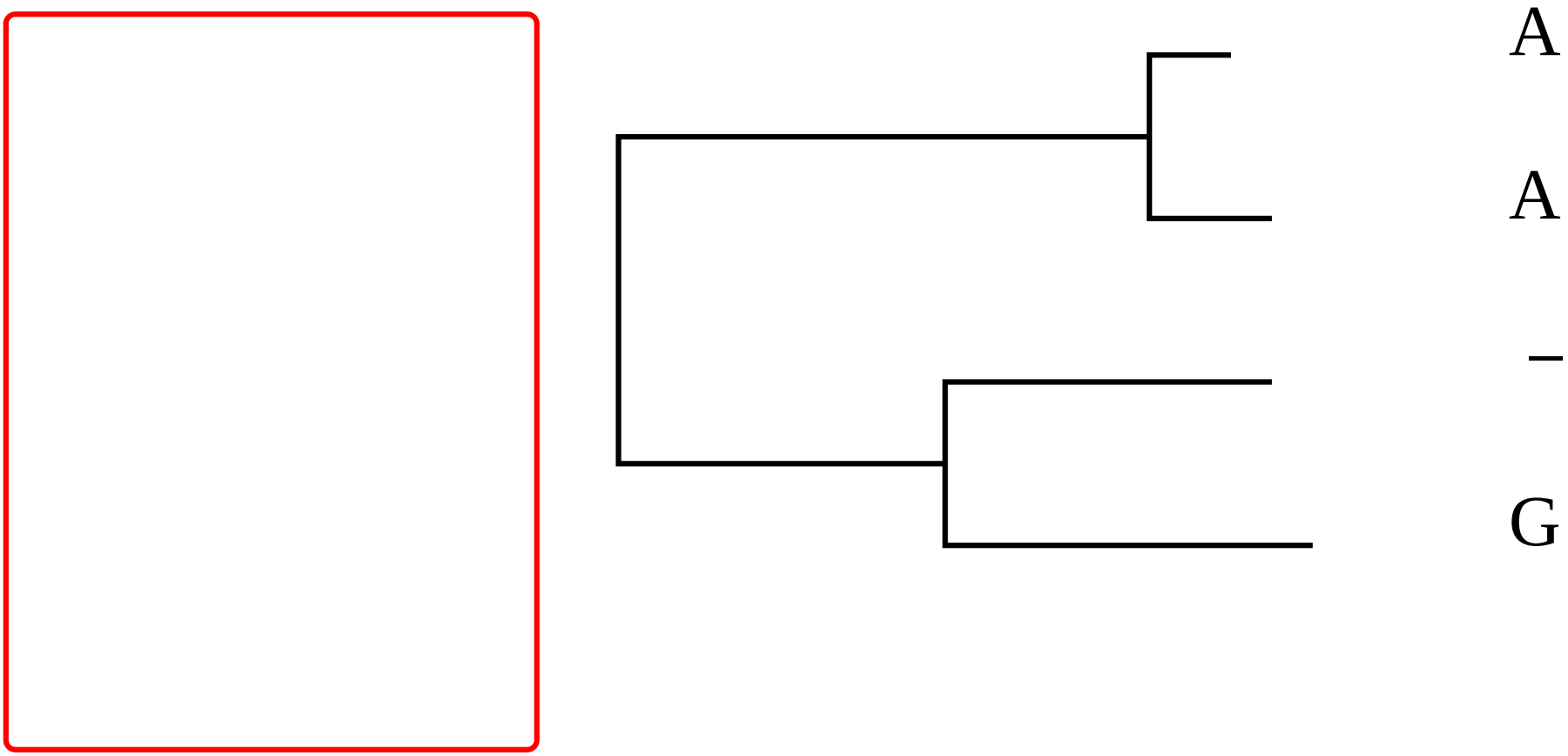
- Modélisation de N dérives ET de K profils biochimiques,
- $K \times (N+1)$ processus de substitutions, $Q_1, ..., Q_{K\times(N+1)}$.

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

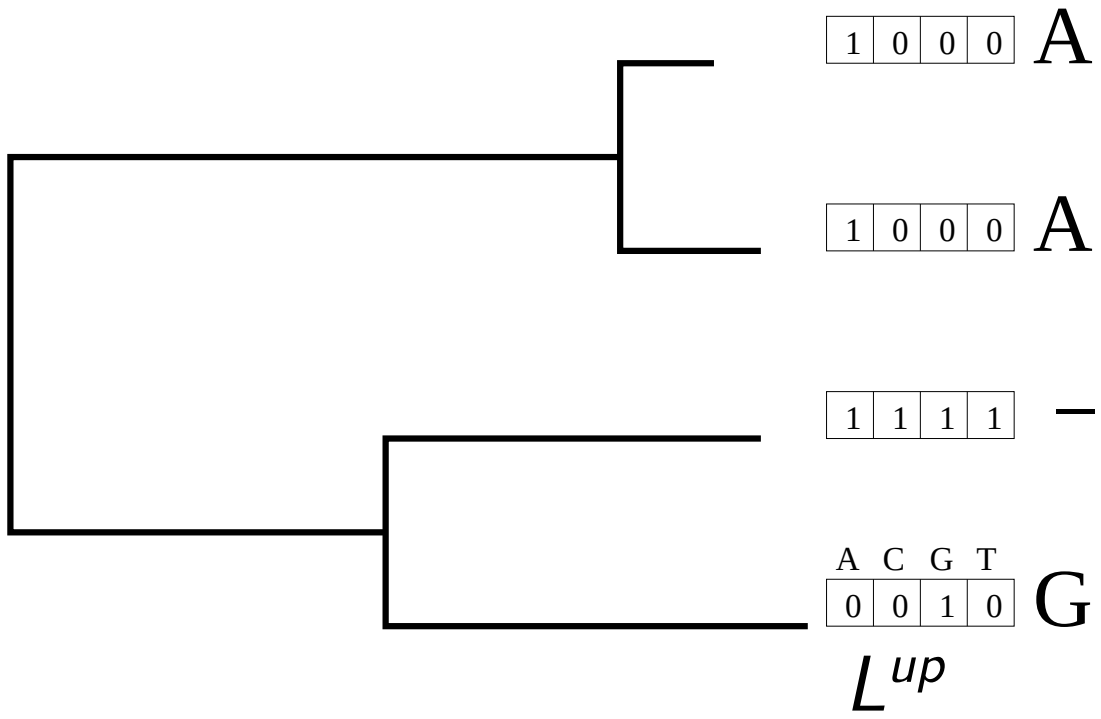
A site- and time-heterogeneous model of amino-acid replacement.

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

Calcul récursif de la vraisemblance (Felsenstein 1981)

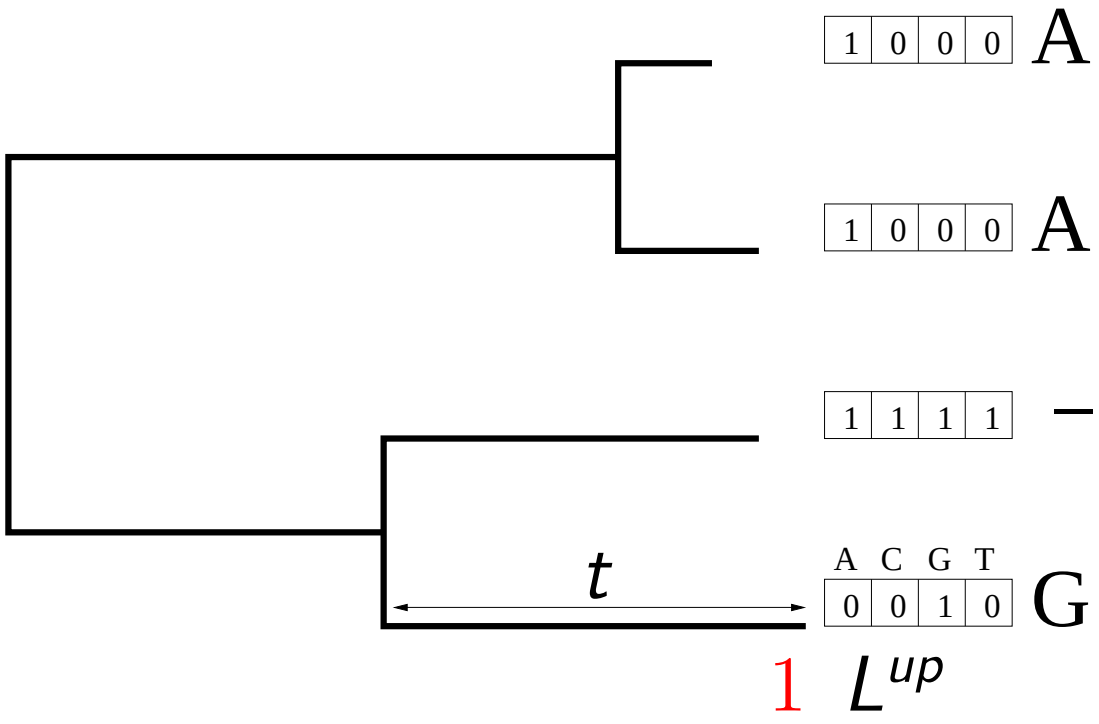


Calcul récursif de la vraisemblance (Felsenstein 1981)



Calcul récursif de la vraisemblance (Felsenstein 1981)

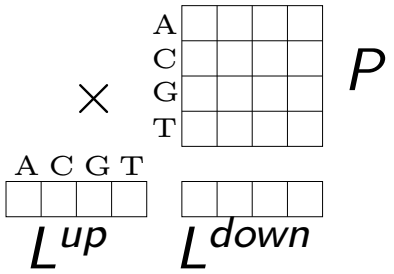
1) $P = e^{tQ}$

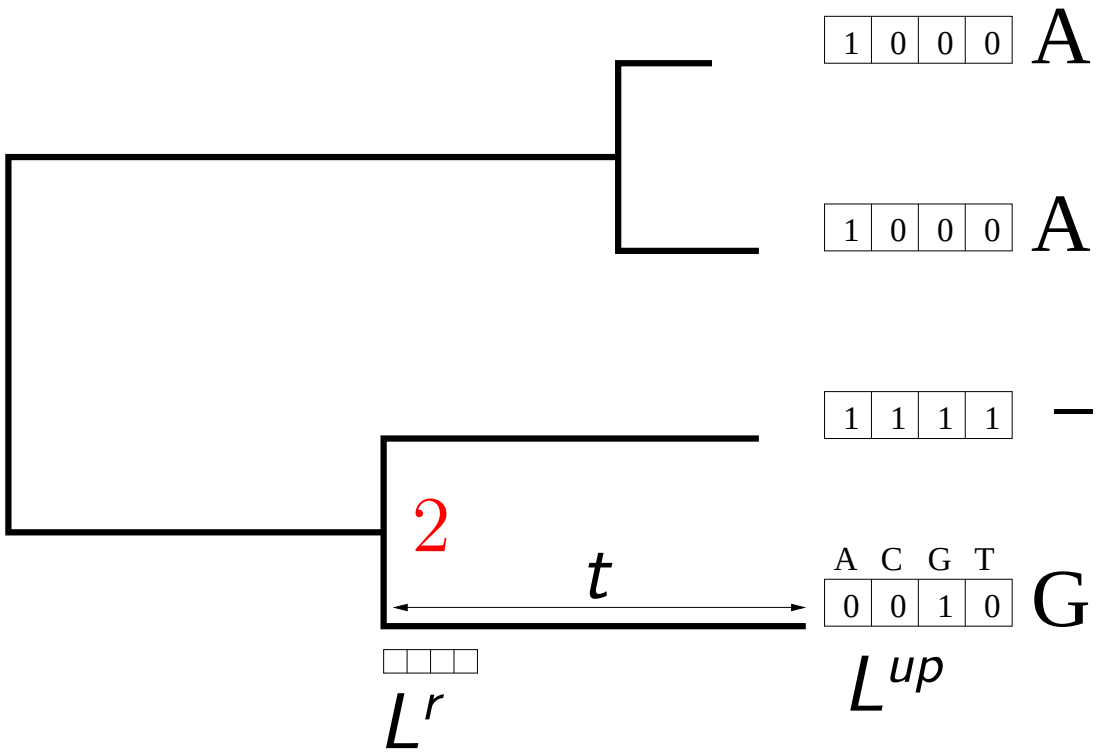


Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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

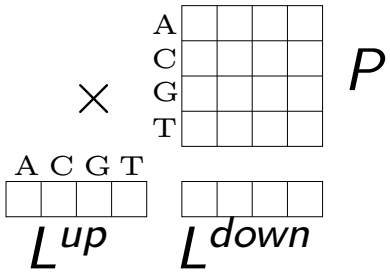


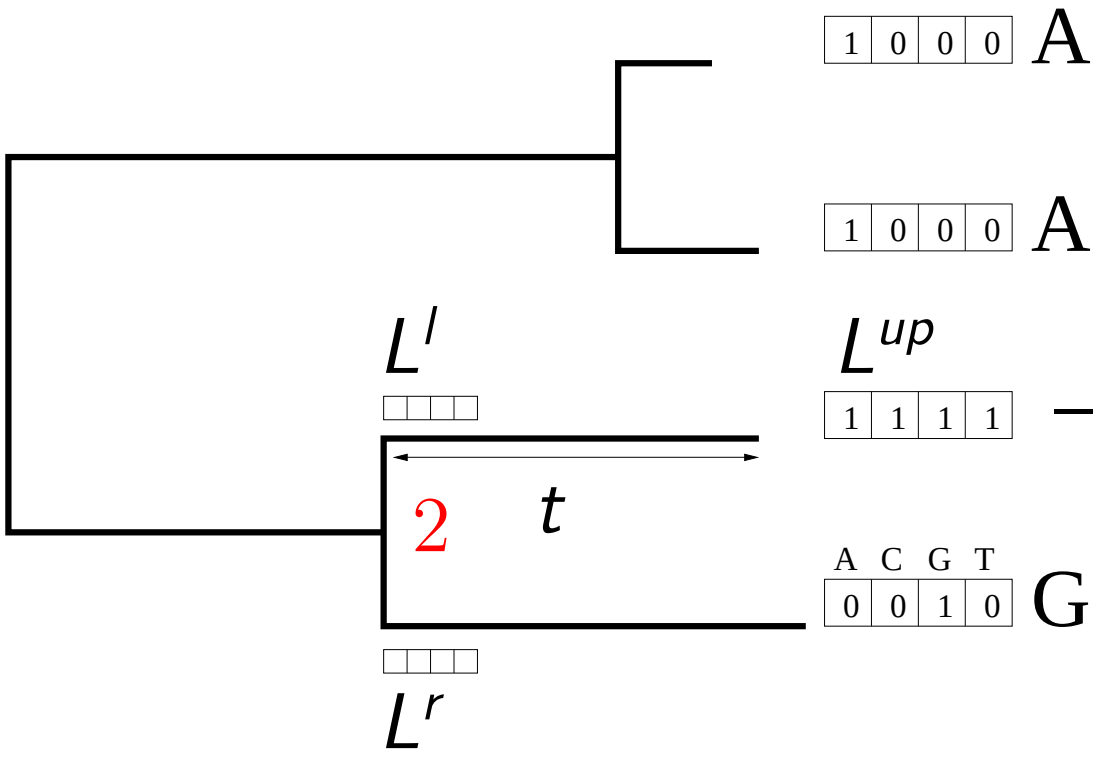


Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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

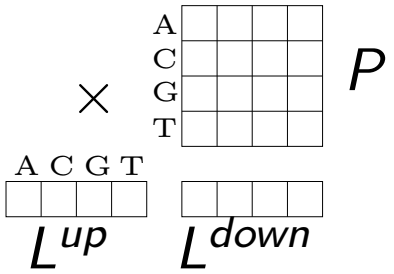




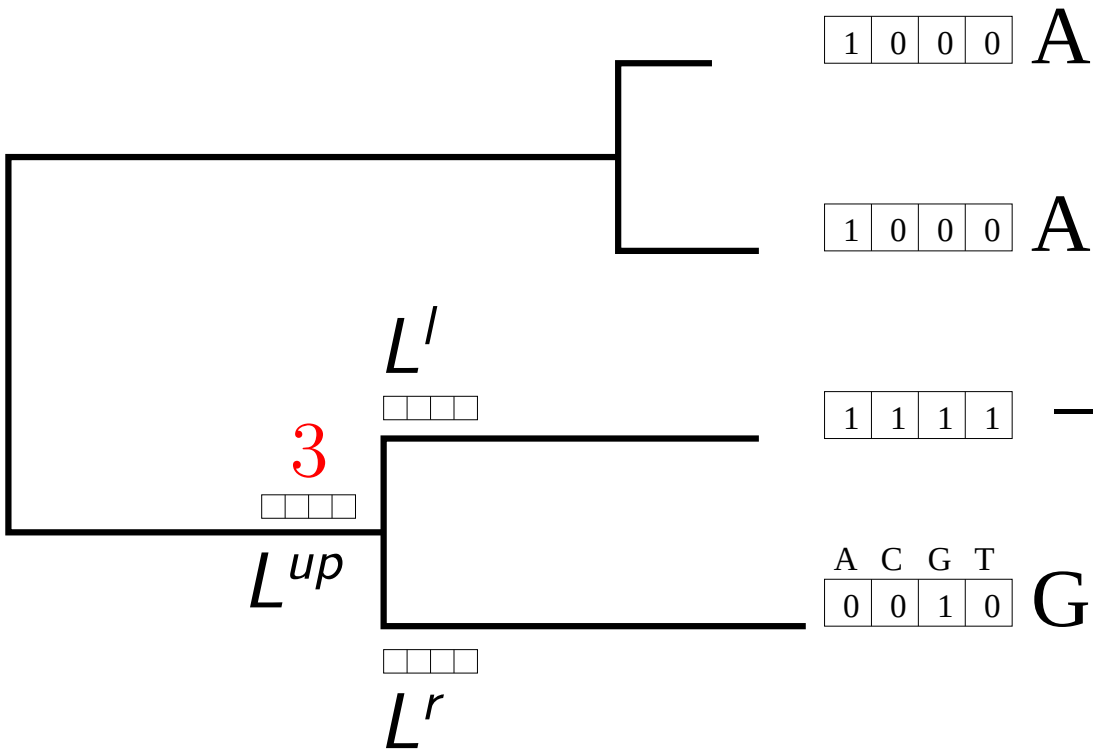
Calcul récursif de la vraisemblance (Felsenstein 1981)

1) $P = e^{tQ}$

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



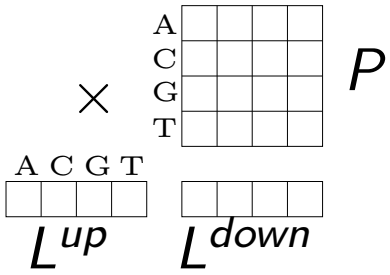
3) $L_i^{up} = L_i^l L_i^r$



Calcul récursif de la vraisemblance (Felsenstein 1981)

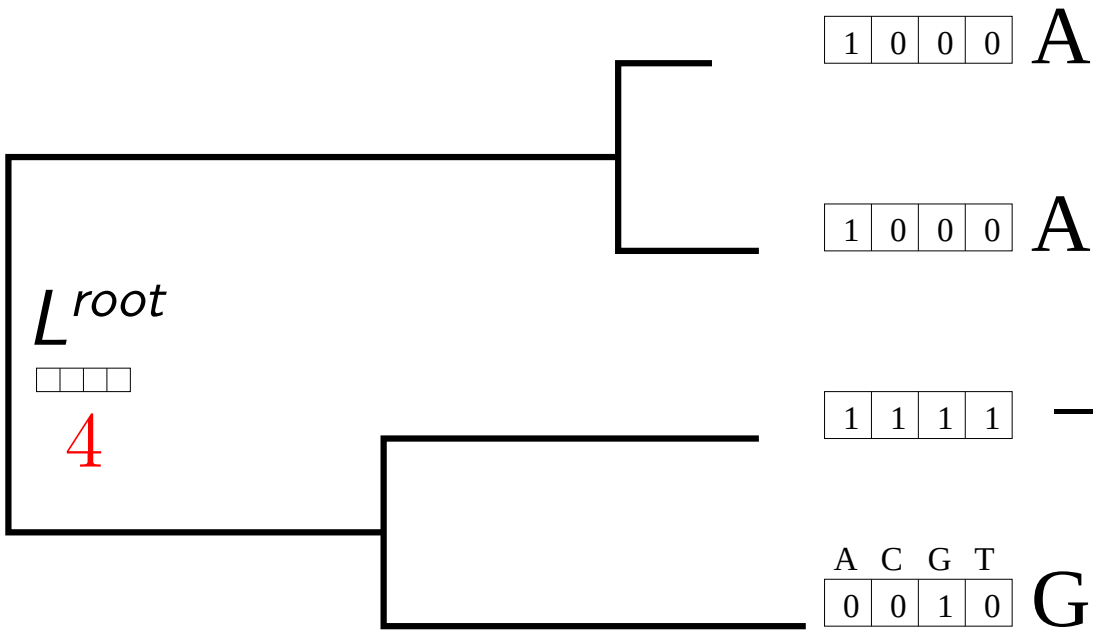
1) $P = e^{tQ}$

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



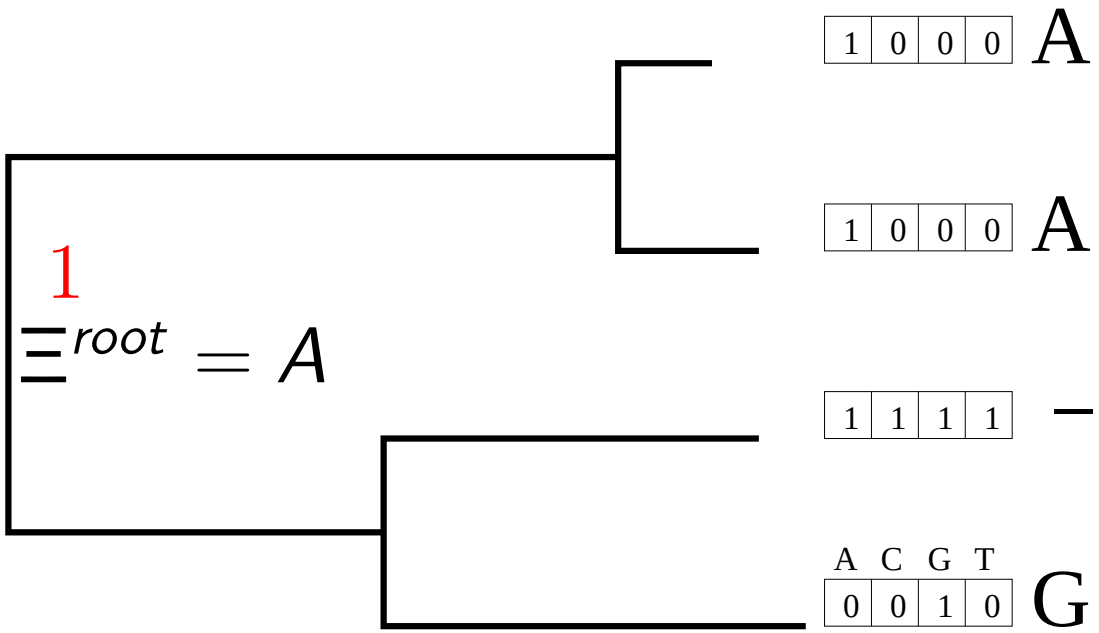
3) $L_i^{up} = L_i^l L_i^r$

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



Construire les séquences ancestrales

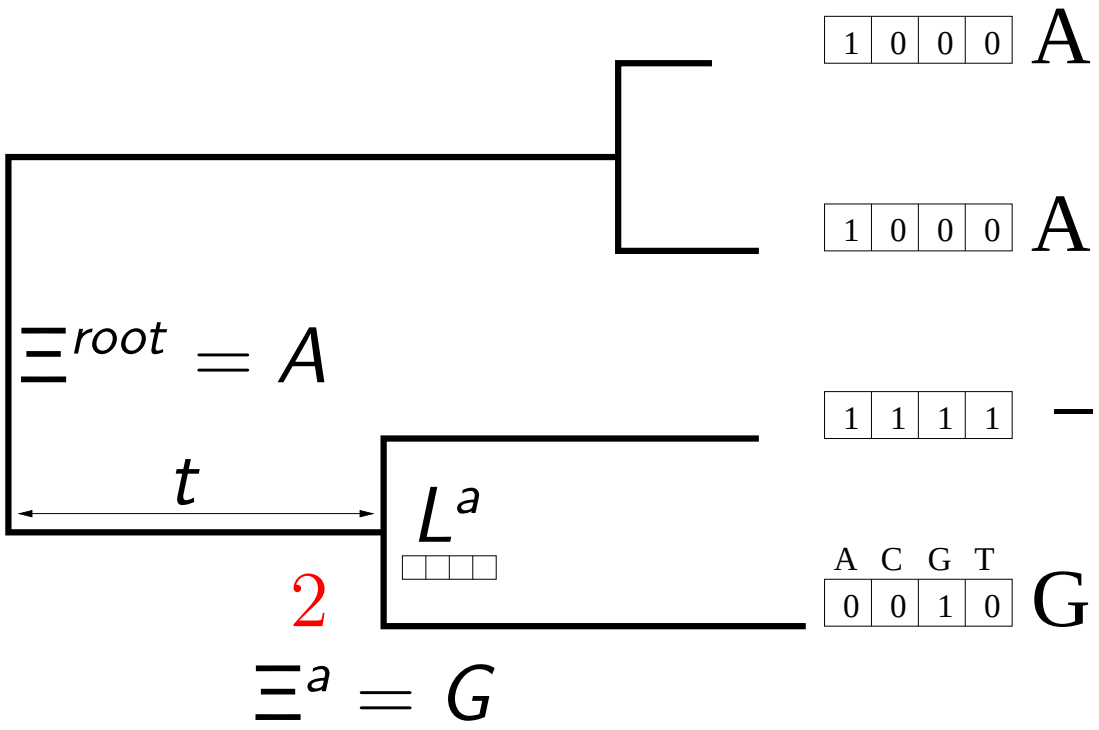
1) $\Xi^{root} \sim \pi_i L_i^{root}$



Construire les séquences ancestrales

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

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

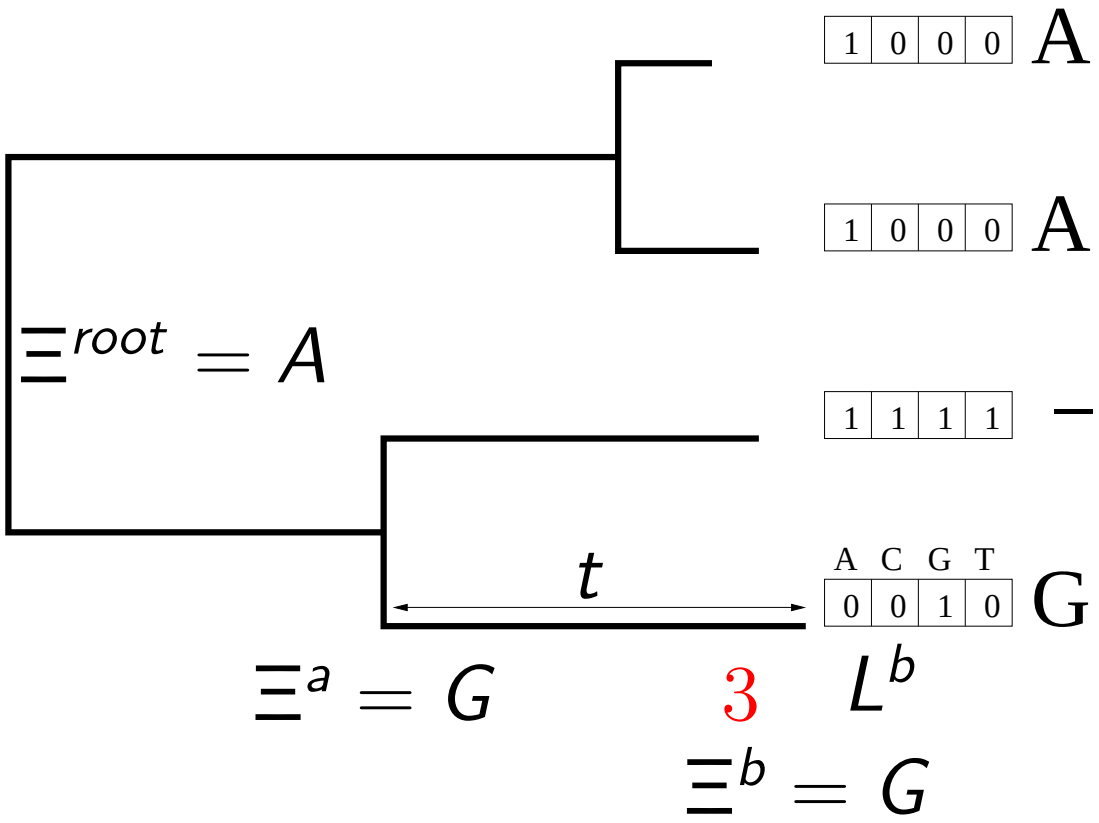


Construire les séquences ancestrales

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

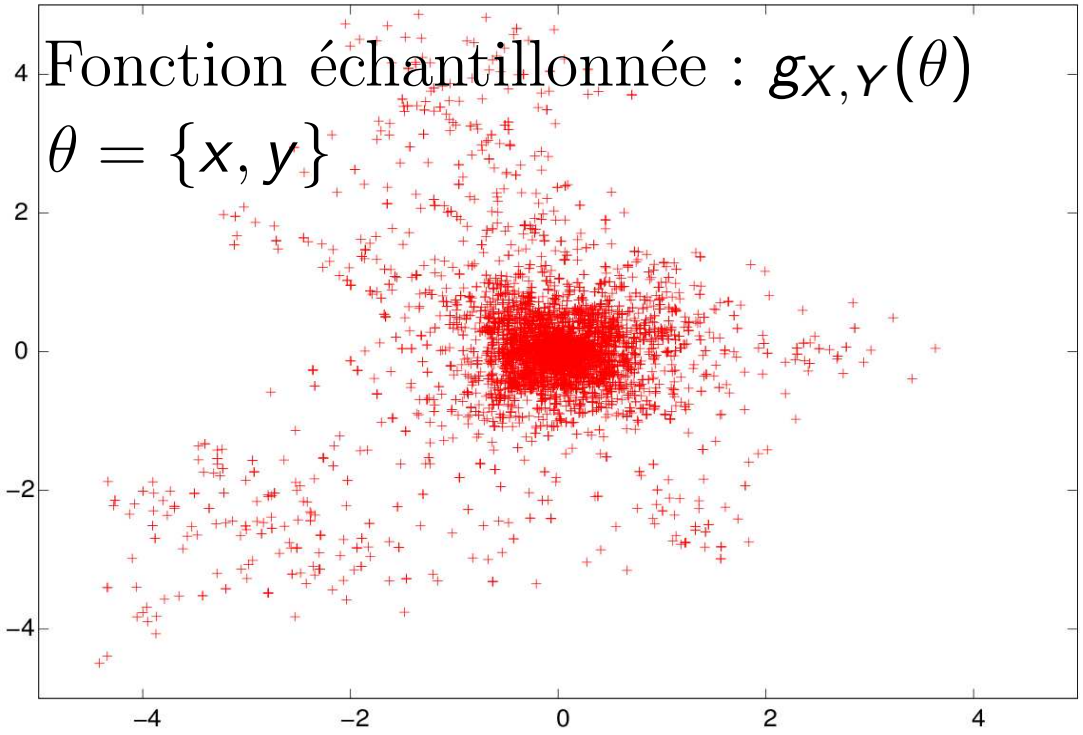
2) $\Xi^{a} \sim P_{\Xi^{root},i} L_i^{a}$

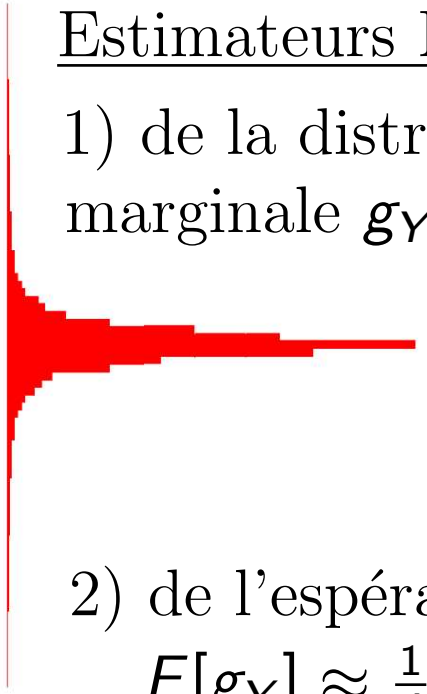
3) $\Xi^{b} \sim P_{\Xi^{a},i} L_i^{b}$



MCMC, Les estimateurs Monte Carlo

- On a obtenu par MCMC une collection de A échantillons $(\theta^a)_{a\in[1..A]}$ tirés de la fonction cible $g_{X,Y}(\theta)$.
- Les approximations de la fonction $g_{X,Y}$ estimables à partir de cette collection sont nommées estimateurs Monte Carlo.





Estimateurs Monte Carlo :

1) de la distribution marginale g_Y :

2) de l'espérance de g_Y :
$E[g_Y] \approx \frac{1}{A}\sum_{a=1}^{A} y^a$