Sur deux problèmes mathématiques de reconstruction phylogénétique

Mikael Falconnet

Institut Fourier - Université de Grenoble

9 juillet 2010

Molecular phylogenetics

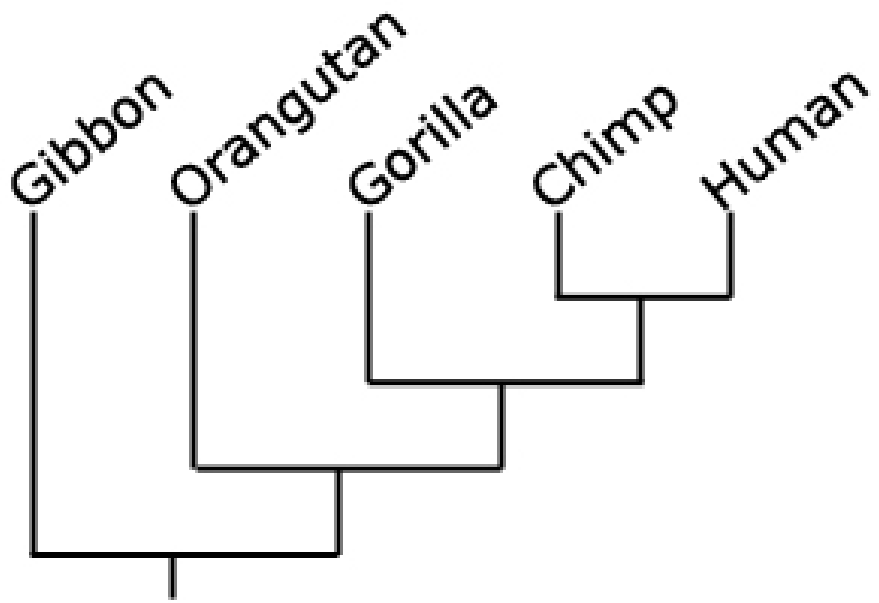
Definition

In biology, phylogenetics is the study of evolutionary relatedness among various groups of organisms (for example, species or populations).

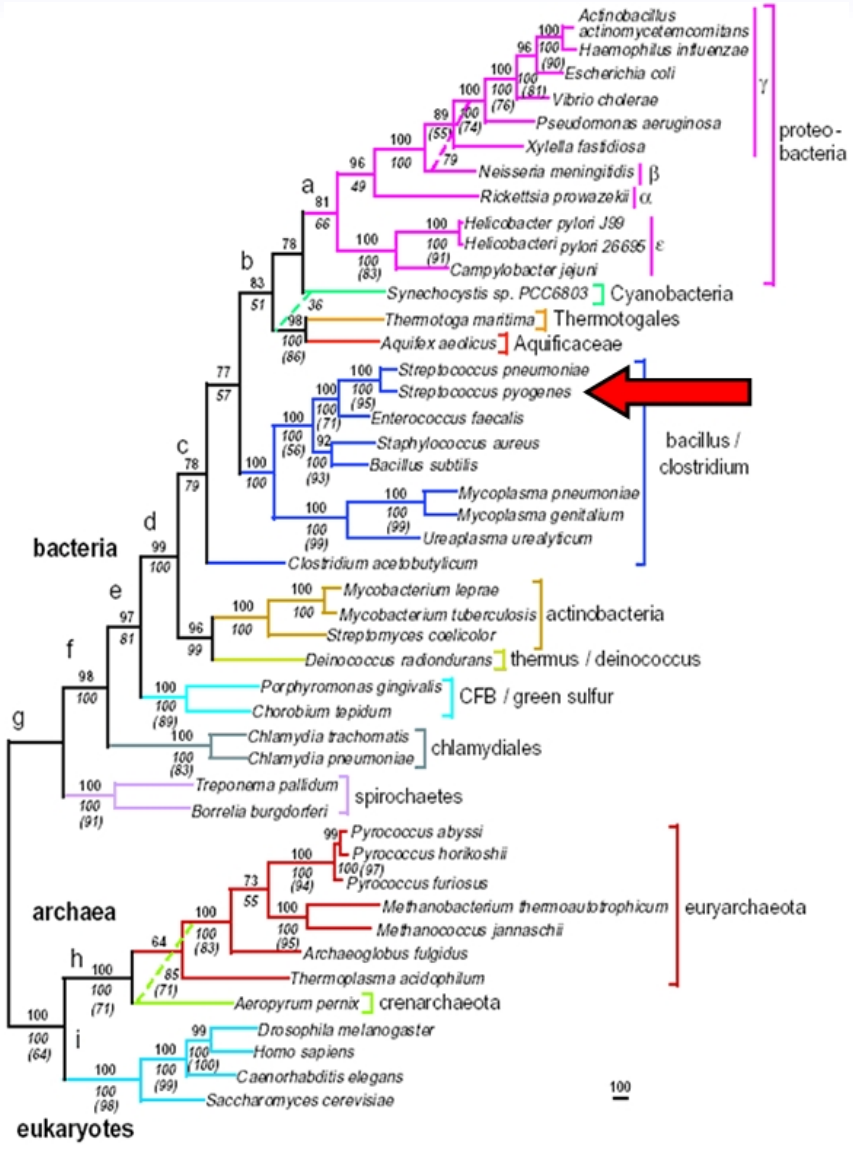
Phylogenetic relationships in the past were reconstructed by looking at **phenotypes**, often anatomical characteristics. Today, molecular data, which includes protein and **DNA sequences**, are used to construct phylogenetic trees.

Phylogenetic tree

Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. This may be visualized in a phylogenetic tree.



Phylogenetic tree

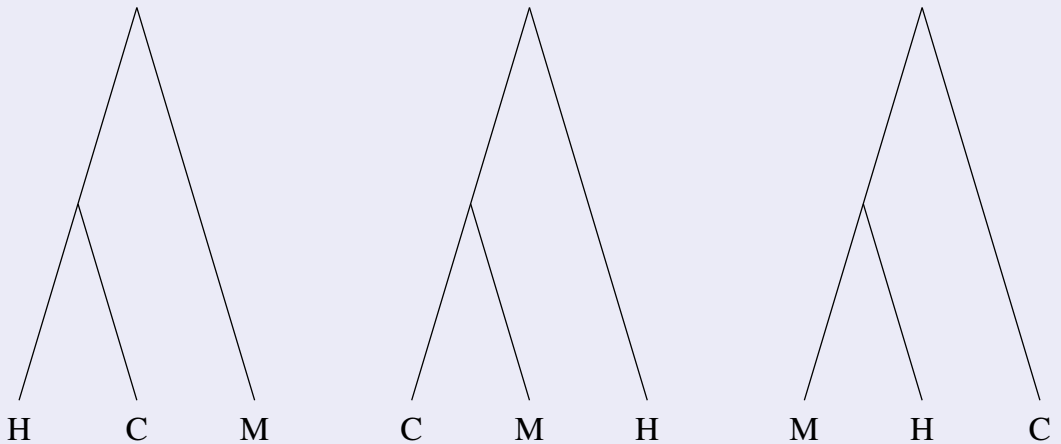


Phylogenetic tree

The **goal** of phylogenetics is to construct **the most probable** phylogenetic tree between **present** organisms.

First problem: topology of the tree

Among the trees below, what is the topology of the phylogenetic tree between the three species: Human, Chimpanzee and Macaque?

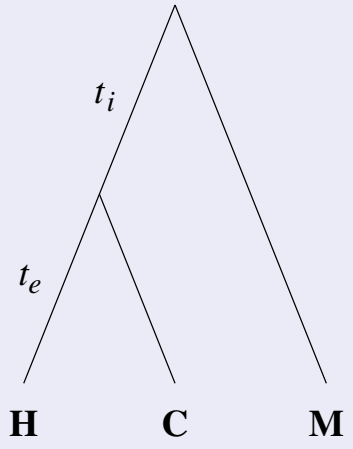


Phylogenetic tree

Assume that one knows the topology of the tree, there is a another question.

Second problem: branch lengthes

What are the branch lengthes (t_i, t_e) in the phylogenetic tree of Human, Chimpanzee, Macaque?



Methods

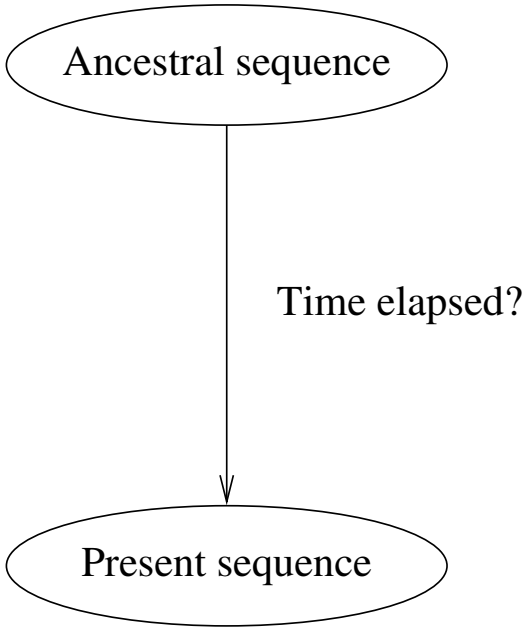
Molecular phylogenetics methods rely on a defined **mutation model** encoding hypothesis about the substitutions, insertions, deletions, etc., of various sites along the DNA sequences being studied.

A **distance-matrix** method is one of the molecular phylogenetic methods, and requires measures of **"genetic distance"** between **DNA sequences**, under a given **mutation model**.

Recent advances in molecular phylogenetics are based on **Bayesian** inference principles.

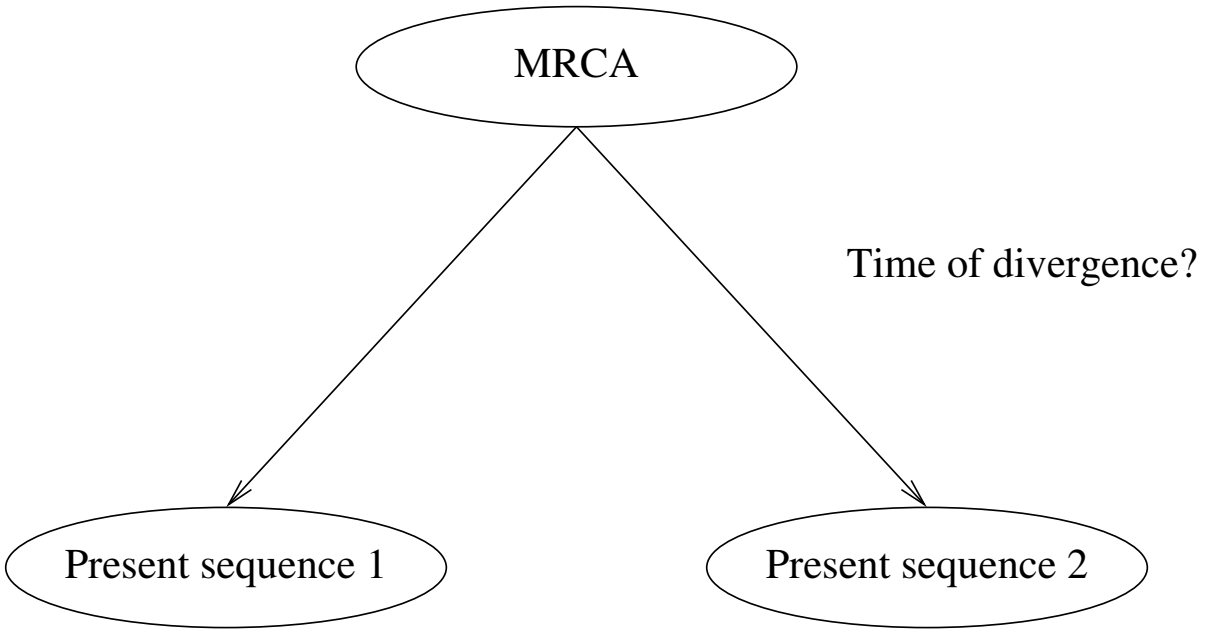
First problem

Assume that a present DNA sequence is issued from an ancestral DNA sequence. We want to estimate the **time elapsed** between these two sequences.



Second problem

Of course, we never have the ancestral DNA sequence, and we have to deal with present DNA sequences only. Assume that two present DNA sequences are issued from a common ancestral DNA sequence. We want to estimate the **time of divergence** between these two sequences.



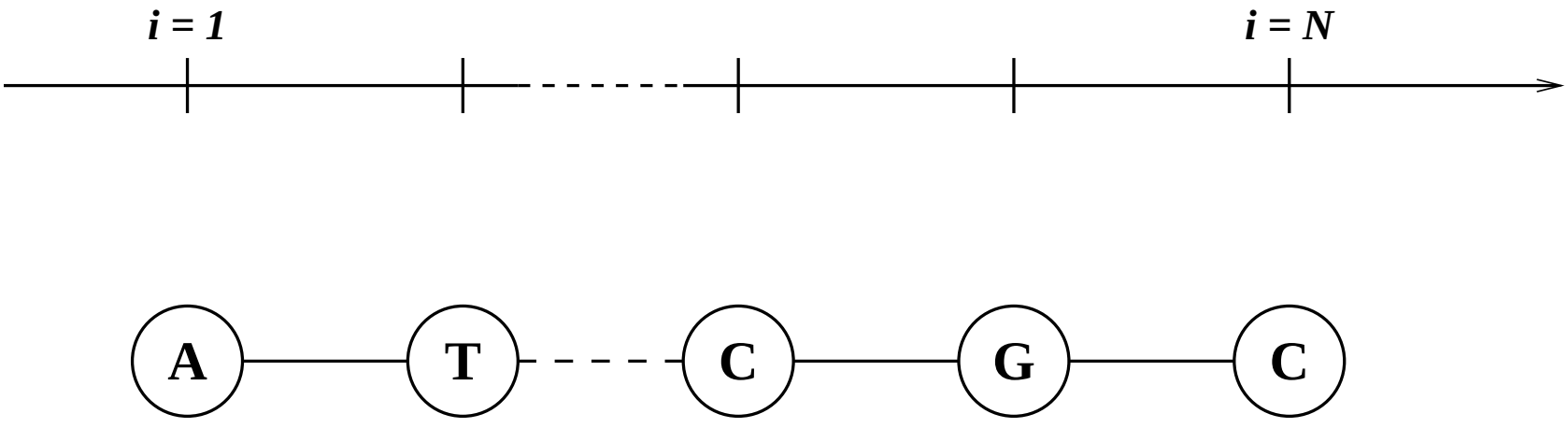
The Jukes-Cantor model

Definition

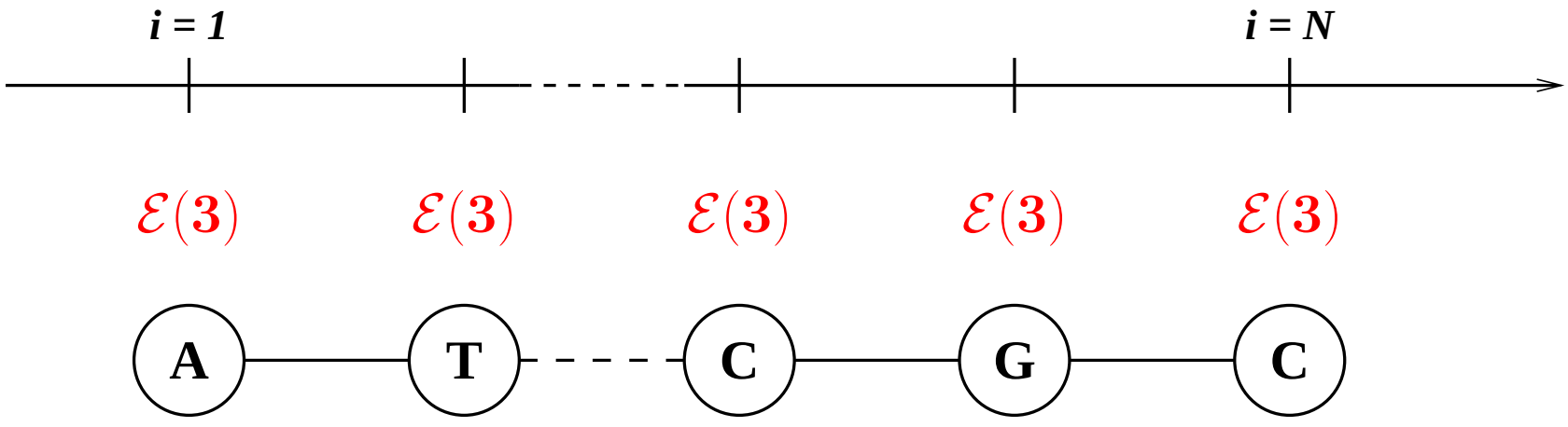
The Jukes-Cantor model is a Markov process $(X_{1:N}(t))_{t\geqslant 0}$ *on* $\mathscr{A}^N$, *where* $\mathscr{A} = \{A, T, C, G\}$ *and* $N \geqslant 1$. *Each site evolves* ***independently*** *from the others with the following infinitesimal generator*

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{pmatrix} \end{array}.$$

Heuristics of JC69



Heuristics of JC69



Why an influence of the neighborood?

CpG sites are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length.

Cytosines in CpG dinucleotides are **methylated** by DNA methyltransferases in many eukaryotic organisms to form 5-methylcytosine. In mammals, 70 per cent to 80 per cent of CpG cytosines are methylated.

Methylated CpGs possess **higher rates of mutations** than non methylated CpGs, and we want to take into account this phenomenon.

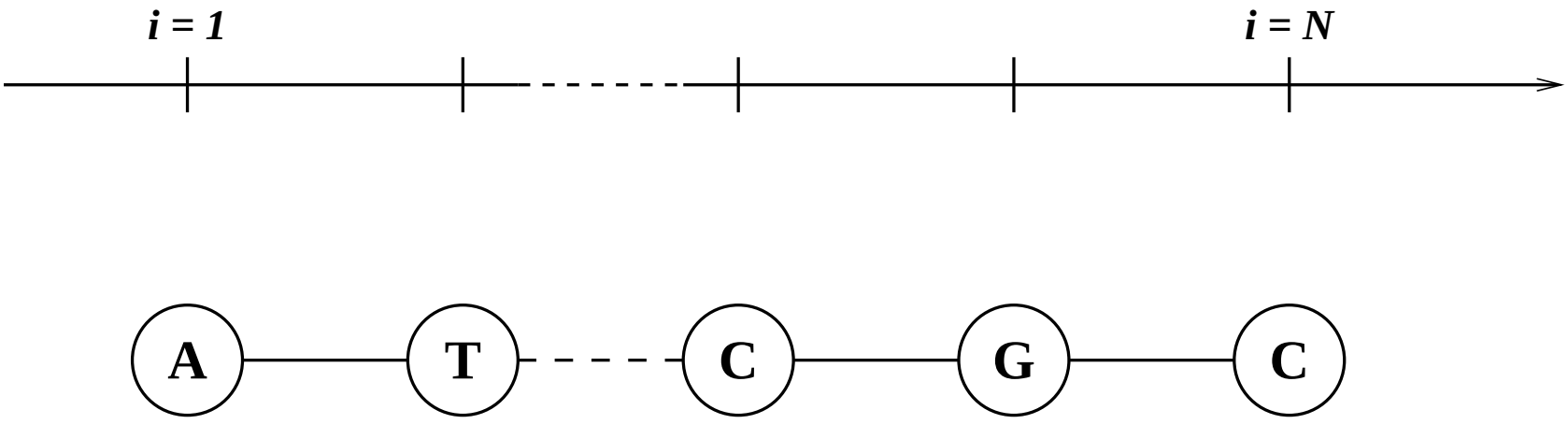
Because of the existence of **CpG islands** in mammalian genomes, that is, a region of DNA with a **higher** concentration of CpG sites, and because these regions are frequently **functional** ones, this phenomenon is important in biology.

Influence of neighborhood in the JC + CpG model

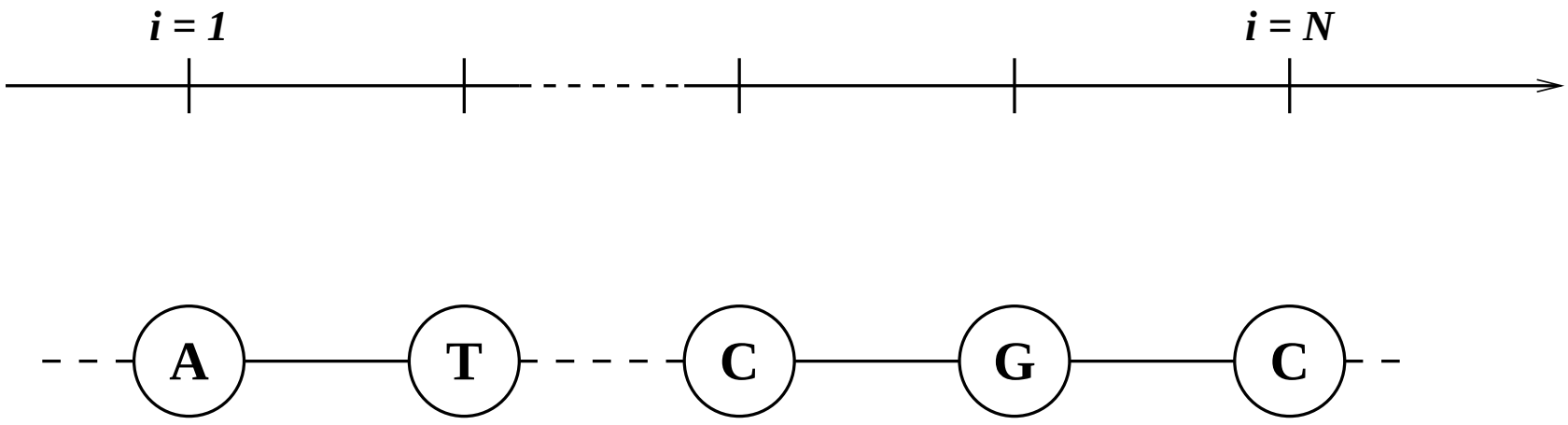
A second mechanism is superimposed to the independent evolution, which describes the substitutions due to the influence of the neighborhood: we assume that the substitution rates of cytosine by thymine and of guanine by adenine in CpG dinucleotides are both increased by an additional nonnegative rate r.

$$CG \xrightarrow{r} TG \quad \text{and} \quad CG \xrightarrow{r} CA.$$

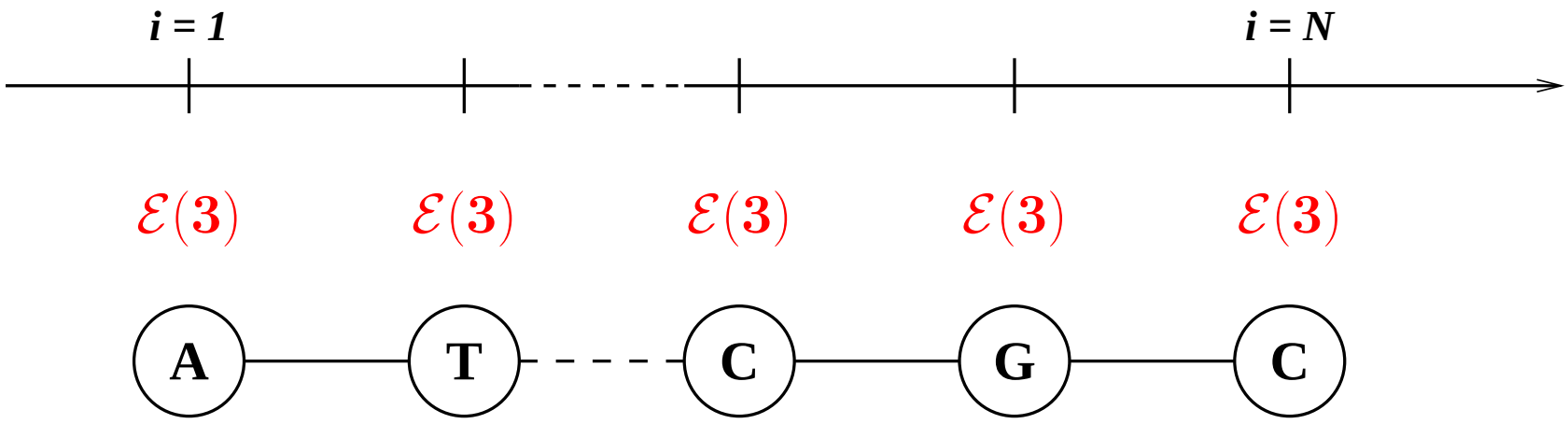
Heuristics of JC+CpG



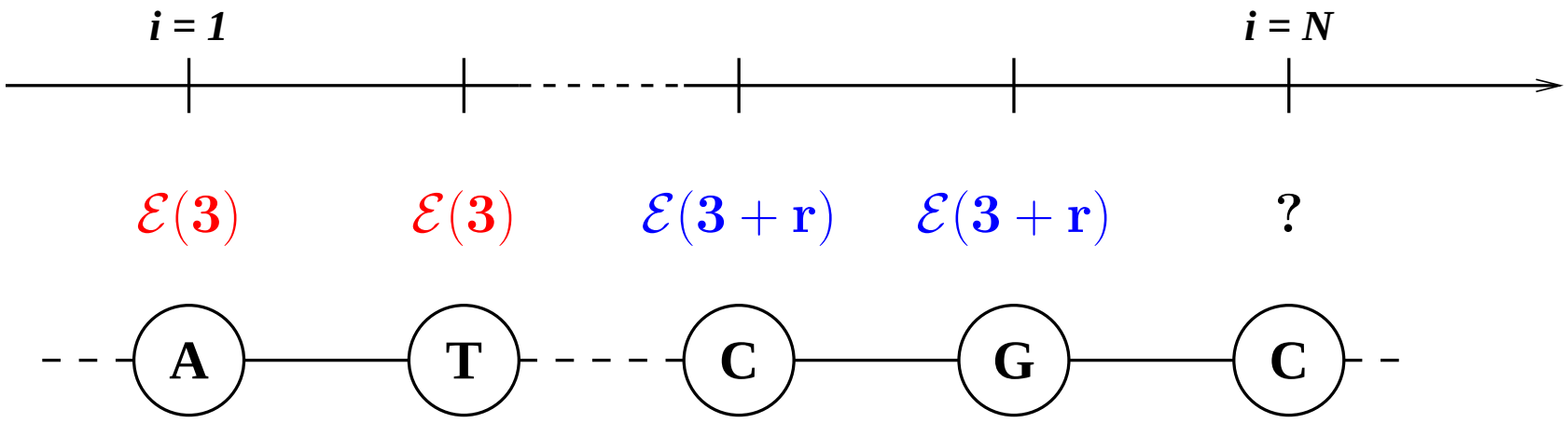
Heuristics of JC+CpG



Heuristics of JC+CpG



Heuristics of JC+CpG



Existence of such a process

Theorem (Bérard, Gouéré and Piau, 2008)

For every probability measure ν on $\mathscr{A}^{\mathbb{Z}}$, there exists a unique Markov process $(X(t))_{t\geqslant 0}$ on $\mathscr{A}^{\mathbb{Z}}$, with initial distribution ν, associated to the transition rates above.

Thus, for every time t, $X(t)$ describes the whole sequence and, for every i in $\mathbb{Z}$, the ith coordinate $X_i(t)$ of $X(t)$ is the random value of the nucleotide at site i and time t.

Ergodicity of the process

Theorem (cont'd)

The process $(X(t))_{t\geqslant 0}$ is ergodic, its unique stationary distribution π on $\mathscr{A}^{\mathbb{Z}}$ is invariant and ergodic with respect to the translations of $\mathbb{Z}$, and π puts a positive mass on every finite word $w = (w_i)_{0\leqslant i\leqslant \ell}$ written in the alphabet $\mathscr{A}$.

2-dependence

Theorem (cont'd)

There exists an i.i.d. sequence $(\xi_i)_{i\in\mathbb{Z}}$ of Poisson processes, and a measurable map Ψ with values in $\mathscr{A}$, such that if one sets

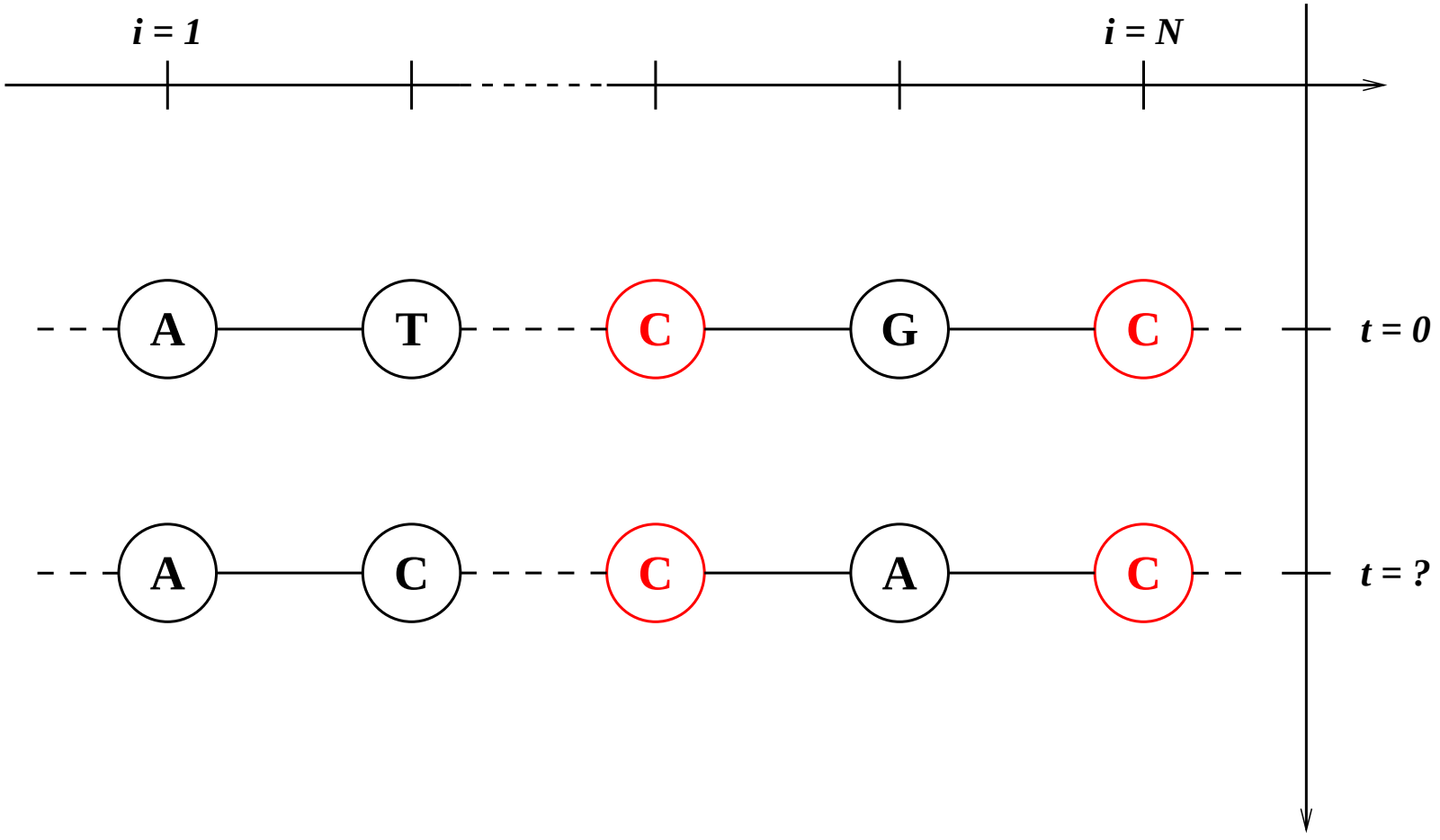
$$\Xi_i = \Psi(\xi_{i-1}, \xi_i, \xi_{i+1})$$

for every site i in $\mathbb{Z}$, then the distribution of $(\Xi_i)_{i\in\mathbb{Z}}$ is π.

In particular, any collections $(\Xi_i)_{i\in I}$ and $(\Xi_i)_{i\in J}$ are independent as soon as the subsets I and J of $\mathbb{Z}$ are such that $|i-j| \geqslant 3$ for every sites i in I and j in J.

We call this property 2-**dependence**.

Idea



Notations and definitions

Definition

Let $(C, C)_{\text{obs}}$ *denote the observed quantity defined as*

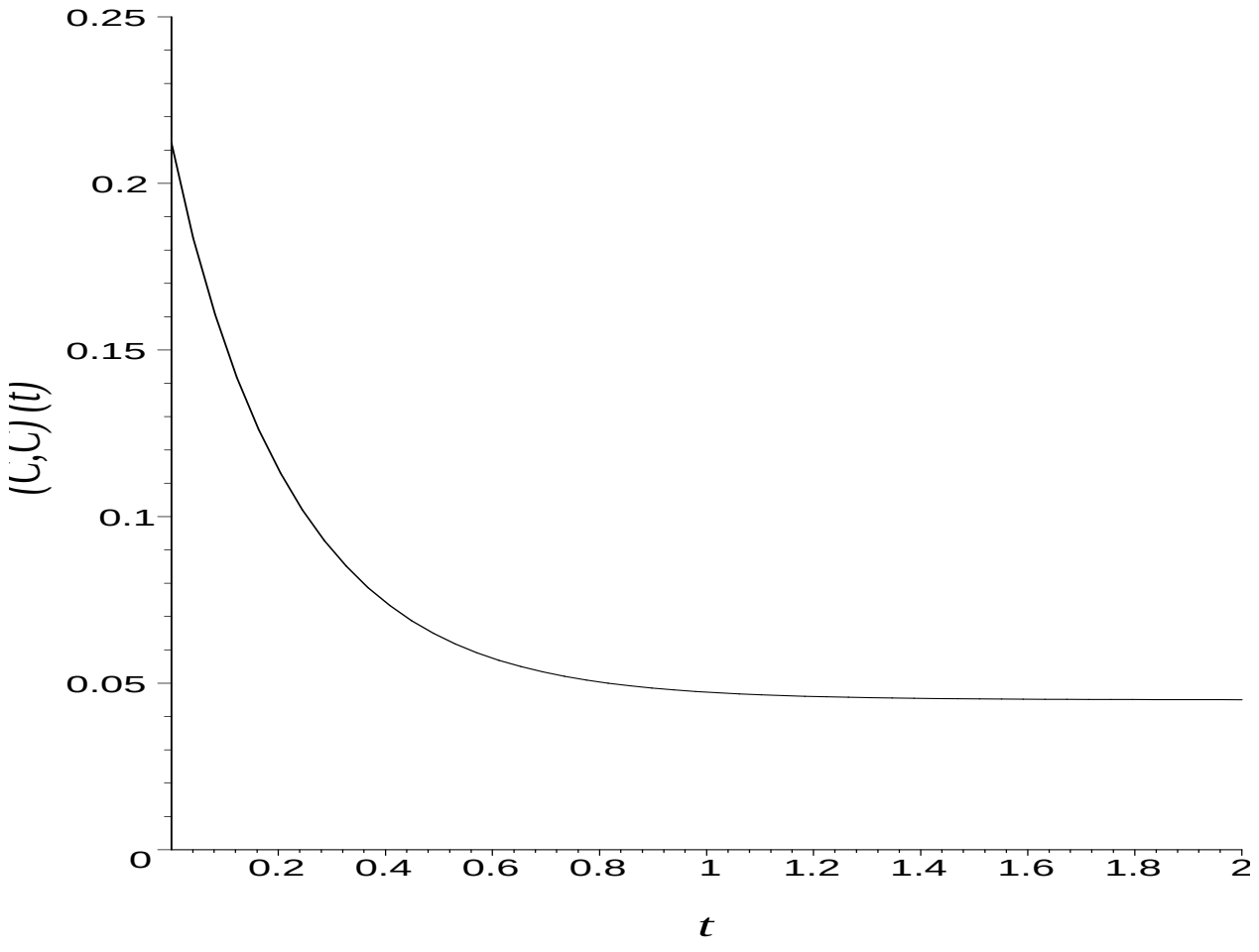
$$(C, C)_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{X_i(t) = C, X_i(0) = C\right\}.$$

Let $(C, C)(t)$ *denote the frequency of sites occupied by* C *at times* 0 *and* t*, that is*

$$(C, C)(t) = \lim_{N \to \infty} (C, C)_{\text{obs}}.$$

The limit above exists thanks to the ergodicity of π with respect to translations.

Representation of $t \mapsto (C, C)(t)$ with $r = 10$



Notations and definitions

Definition

Let T_C denote the estimator of the elapsed time as the solution in τ of the equation

$$(C, C)(\tau) = (C, C)_{\text{obs}}.$$

Let κ^C_{obs} and ν^C_{obs} denote observed quantities, defined as

$$\begin{aligned}
\kappa^C_{\text{obs}} &= 4(C, C)_{\text{obs}} + r(C*, CG)_{\text{obs}} - (C)_{\text{obs}}, \\
\nu^C_{\text{obs}} &= (C, C)_{\text{obs}} + 2(CC, CC)_{\text{obs}} + 2(C * C, C * C)_{\text{obs}} \\
&\quad - 5(C, C)^2_{\text{obs}}.
\end{aligned}$$

Result

Theorem (MF, 2010)

Assume that the ancestral sequence is at stationarity. Then, in the JC+CpG model,

$$\kappa_{\mathrm{obs}}^{C}\sqrt{N/\nu_{\mathrm{obs}}^{C}}(T_C - t) \xrightarrow[N\to+\infty]{d} \mathcal{N}(0,1).$$

An asymptotic confidence interval at level ε for the elapsed time is

$$\left[T_C - \frac{z(\varepsilon)}{\kappa_{\mathrm{obs}}^{C}}\sqrt{\frac{\nu_{\mathrm{obs}}^{C}}{N}}, T_C + \frac{z(\varepsilon)}{\kappa_{\mathrm{obs}}^{C}}\sqrt{\frac{\nu_{\mathrm{obs}}^{C}}{N}}\right],$$

where $z(\varepsilon)$ denotes the unique real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ with Z a standard normal random variable.

Steps of the proof

Step 1

We state a central limit theorem for $(C, C)_{\mathrm{obs}}$

$$\sqrt{N}\left((C, C)_{\mathrm{obs}} - (C, C)(t)\right) \xrightarrow[N \to +\infty]{d} \mathcal{N}(0, \sigma_C^2(t)).$$

Steps of the proof

Step 2

We prove that $t \mapsto (C, C)(t)$ is a diffeomorphism and we use the Delta method to state

$$\sqrt{N}(T_C - t) \xrightarrow[N \to +\infty]{d} \mathcal{N}(0, \sigma_C^2(t)/(C, C)'(t)^2).$$

Steps of the proof

Step 3

Using the fact that

$$\kappa_{\text{obs}}^{C} \xrightarrow[N\to+\infty]{a.s.} -(C,C)'(t),$$

and

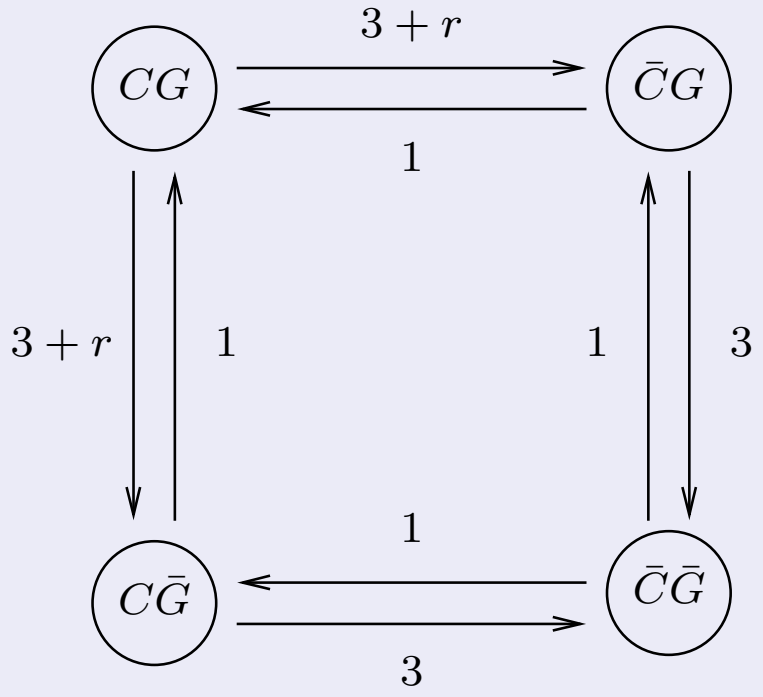
$$\nu_{\text{obs}}^{C} \xrightarrow[N\to+\infty]{a.s.} \sigma_C^2(t),$$

we state the result with Slutsky's lemma.

Step 2

Dynamics

In the JC+CpG model, dinucleotides coded as $\{C, \bar{C}\} \times \{G, \bar{G}\}$ have autonomous evolution whose dynamics can be represented with the following graph



Step 2

Consider the positive real numbers u, u_+ and u_- defined as

$$u = \sqrt{4 + 2r + r^2}, \quad u_+ = 6 + r + u, \qquad u_- = 6 + r - u.$$

Proposition

In the stationary regime,

$$(C, C)(t) = c_0 \mathrm{e}^{-4t} + c_+ \mathrm{e}^{-u_+ t} + c_- \mathrm{e}^{-u_- t} + (C)_*^2,$$

with

$$c_0 = \frac{3 + r}{2(16 + 5r)},$$

and,

$$c_\pm = \frac{3 + r}{4u(16 + 5r)^2} \left(u(16 + 3r) \mp (32 + 14r + 3r^2) \right).$$

Step 2

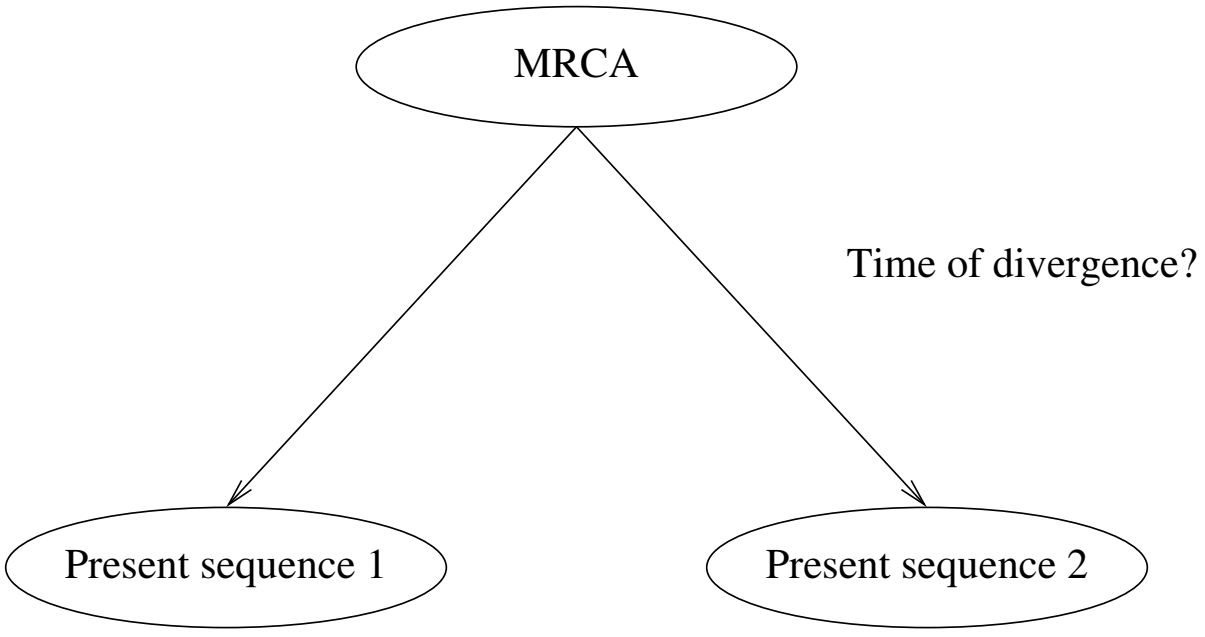
For every positive r, the parameters $c_{\pm}$ and c_0, are positive. This proves the following corollary.

Corollary

In the JC+CpG model, the function $t \mapsto (C, C)(t)$ *is a decreasing diffeomorphism from* $[0, +\infty)$ *to* $((C)^2_*, (C)_*]$.

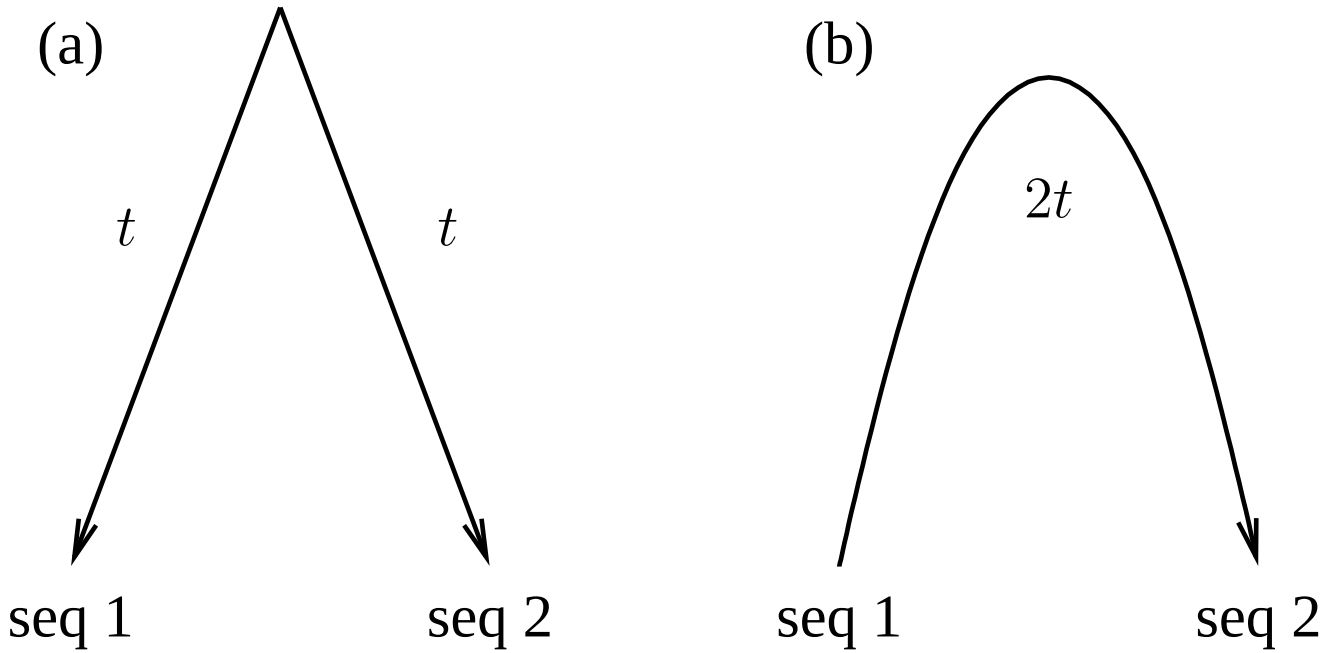
Second problem

Of course, we never have the ancestral DNA sequence, and we have to deal with present DNA sequences only. Assume that two present DNA sequences are issued from a common ancestral DNA sequence. We want to estimate the **time of divergence** between these two sequences.



Reversibility

Although the JC+CpG model is **not reversible**, the dynamics of dinucleotides encoded as $\{C, \bar{C}\} \times \{G, \bar{G}\}$ with respect to this model is **reversible**.



Reversibility

Proposition

For every positive t, we have

$$[C, C](t) = (C, C)(2t).$$

where $[C, C](t)$ denote the frequency of sites occupied by C in the left sequence (denoted by X^1) and the right sequence (denoted by X^2) at time t.
However, as soon as $r > 0$,

$$[A, A](t) \neq (A, A)(2t).$$

Bayesian philosophy

The Bayesian inference in phylogenetics is based on a quantity called a **posterior probability distribution** on trees.

Bayes's Theorem

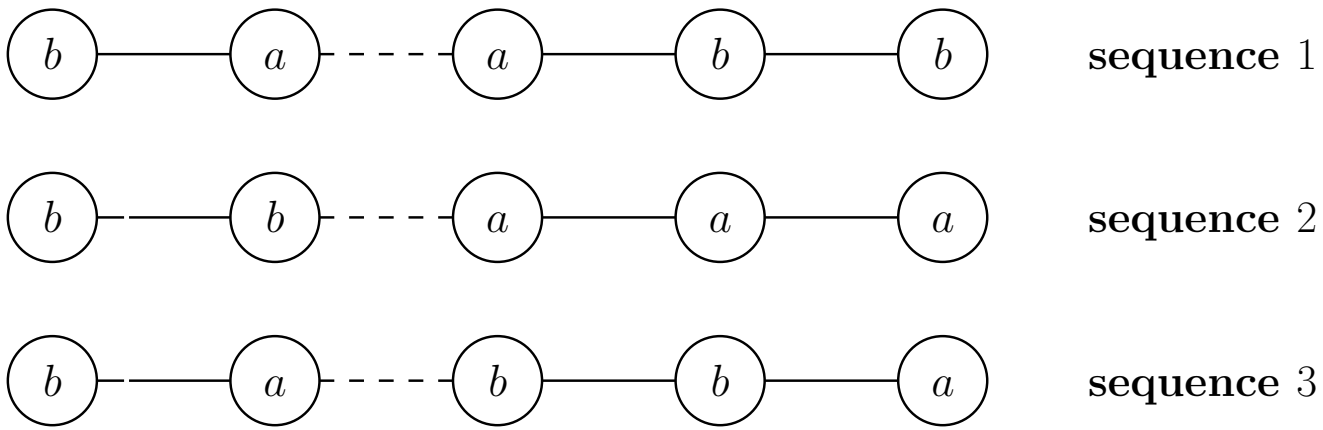
$$\mathbb{P}(\text{Tree}|\text{Data}) = \frac{\mathbb{P}(\text{Data}|\text{Tree}) \times \mathbb{P}(\text{Tree})}{\mathbb{P}(\text{Data})},$$

where

- $\mathbb{P}(\text{Tree})$ denotes the **prior probability** of a phylogeny,
- $\mathbb{P}(\text{Data}|\text{Tree})$ denotes the **likelihood**.
- $\mathbb{P}(\text{Data})$ is a normalization constant and is not needed.

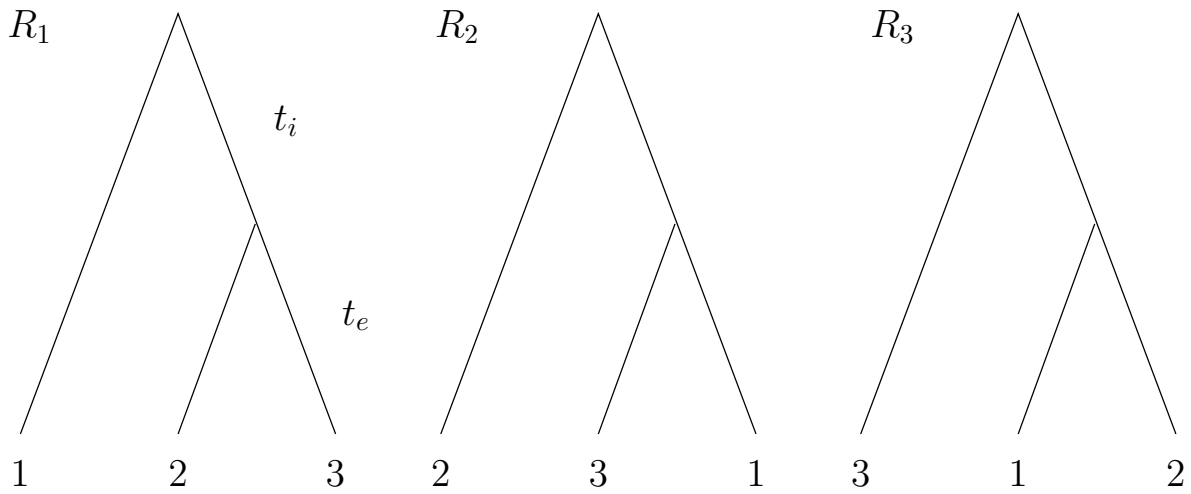
The objects

Let $\tau = \{1, 2, 3\}$ denote a set of three sequences of $\{a, b\}^n$.



The objects

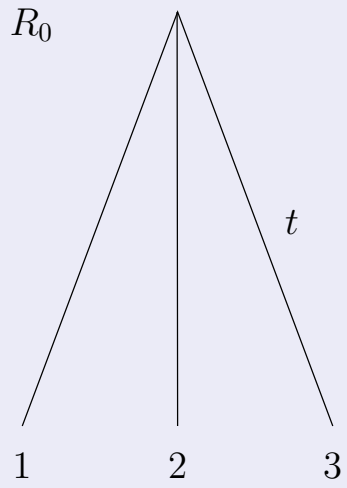
A **resolved** phylogeny on τ is one of the trees below.



The data

Hypothesis

Assume that the sequences of length n are generated by the star tree R_0 on three taxa below with strictly positive edge t, under the **2-state symmetric Markov process**.



The paradox

Theorem (Steel & Matsen, 2007)

*For the uniform prior distribution on (R_1, R_2, R_3), for a specific class of branch length priors (so called **tame** priors), for every fixed $\varepsilon > 0$, the posterior probability of any of the possible trees stays above $1 - \varepsilon$ with non-vanishing probability when the length of the sequence tends to infinity.*

Theorem (MF,2010+)

*Steel and Matsen's conclusion holds for a wider class of branch length priors which we called **tempered**.*

Tame

Definition

*A prior distribution on $\mathfrak{T} = (T_e, T_i)$ is **tame** if it has a smooth joint probability density function which is bounded and everywhere non zero.*

Example

The distribution of a pair of independent exponential random variables is tame.

Tempered

Proposition

Assume that the distribution of $\mathfrak{T} = (T_e, T_i)$ is tame, then this distribution is tempered.

Proposition

Assume that T_e is a continuous random variable, with exponential law. Assume that T_i is a random variable independent of T_e. Then the following holds.

- *If the distribution of T_i is uniform on $[0, \theta]$, with $\theta > 0$, the distribution of $\mathfrak{T} = (T_e, T_i)$ is tempered but not tame.*
- *If the distribution of T_i has density $\theta t_i^{\theta-1}$ on the interval $[0, 1]$, for a given $\theta \in (0, 1)$, the distribution of $\mathfrak{T} = (T_e, T_i)$ is tempered but not tame.*
- *If the distribution of T_i has density $4t_i \log(1/t_i)$ on the interval $[0, 1]$, the distribution of $\mathfrak{T} = (T_e, T_i)$ is not tempered.*

Tempered

Proposition

Assume the following:

(i) *The random variable* T_{i} *is discrete and such that, for every* $n \geqslant 1$,

$$\mathbb{P}(T_{\mathrm{i}} = t_n) = p_n,$$

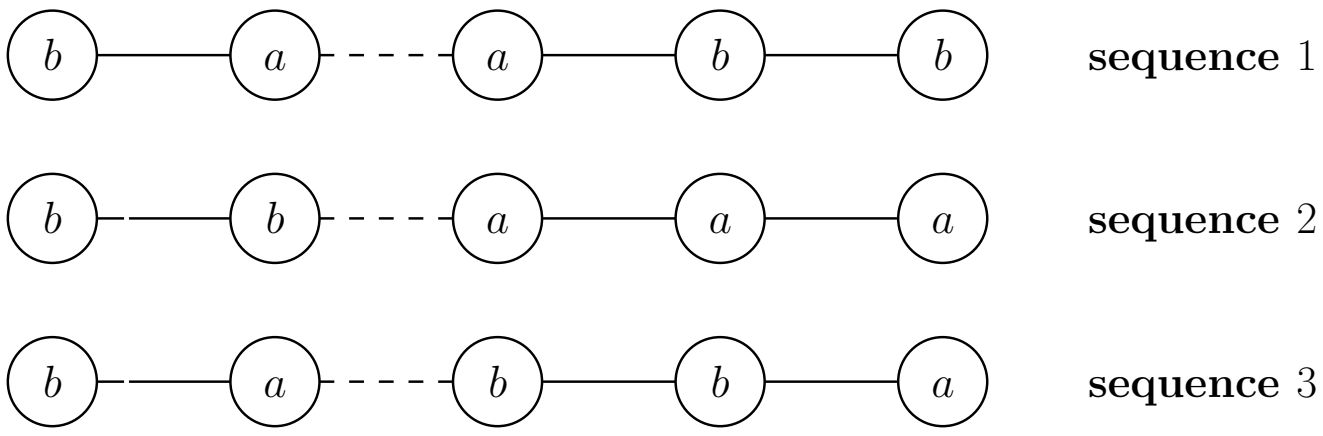
for suitable (t_n) *and* (p_n).

(ii) *The random variable* T_{e} *is continuous, independent of* T_{i}, *with exponential law of parameter* 4, *that is, with density* $4\,\mathrm{e}^{-4t}$ *on* $t \geqslant 0$ *with respect to the Lebesgue measure.*

Then, the distribution of $\mathfrak{T} = (T_{\mathrm{e}}, T_{\mathrm{i}})$ *is not tame but it is tempered.*

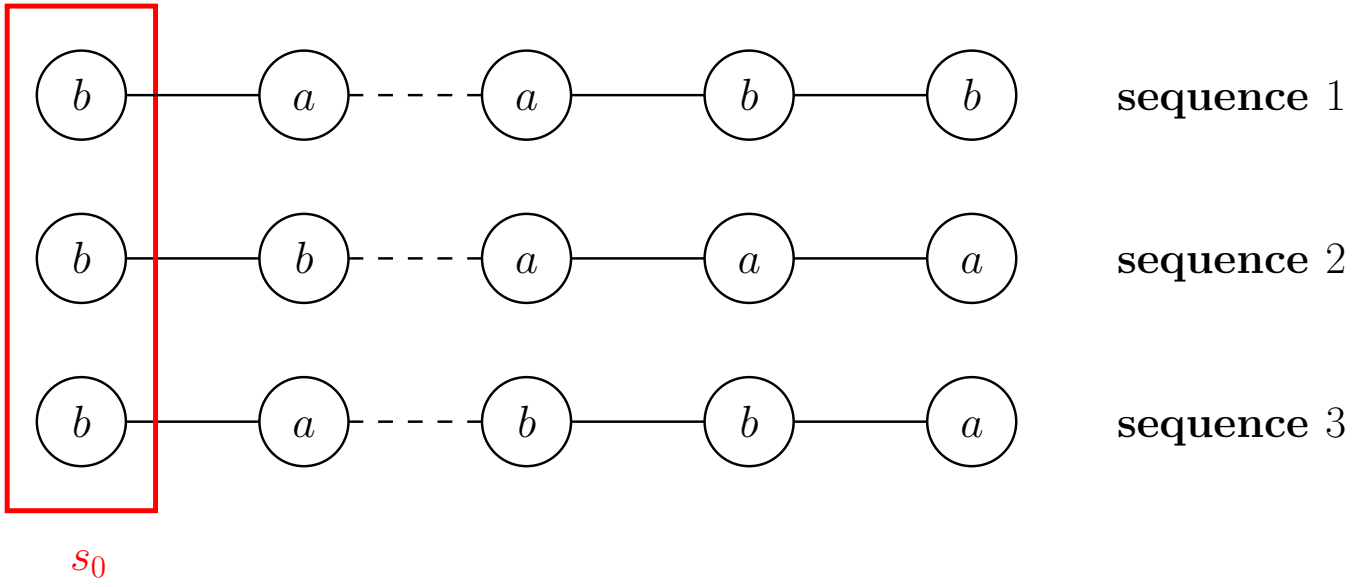
Site patterns

Four site patterns can occur on τ.



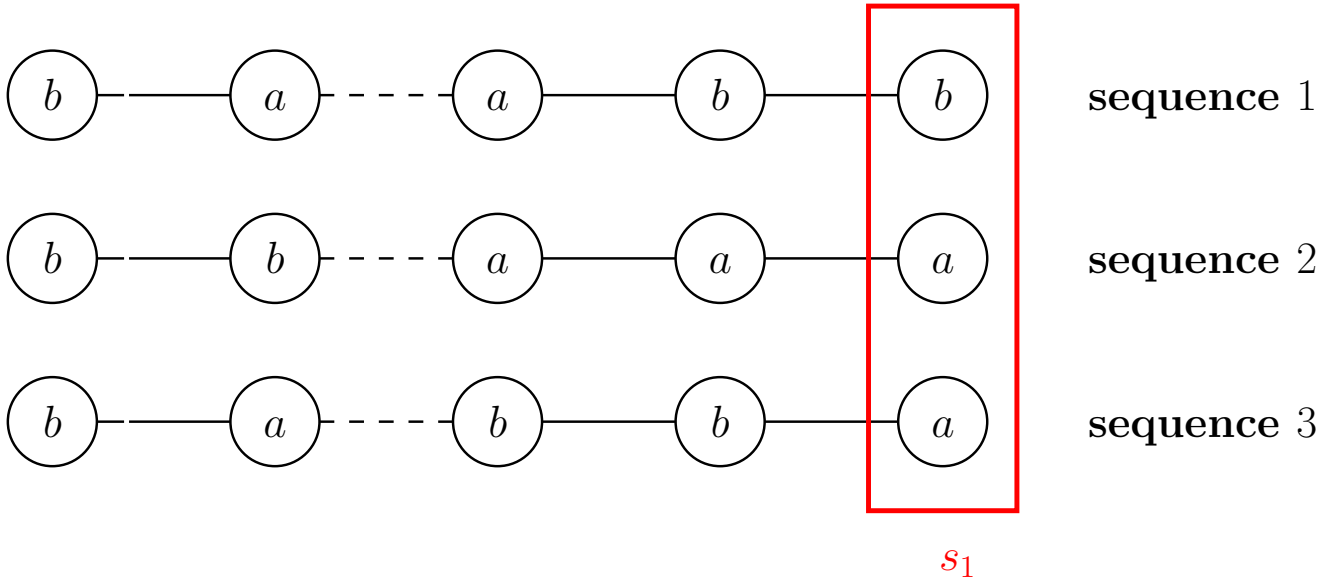
Site patterns

Four site patterns can occur on τ.



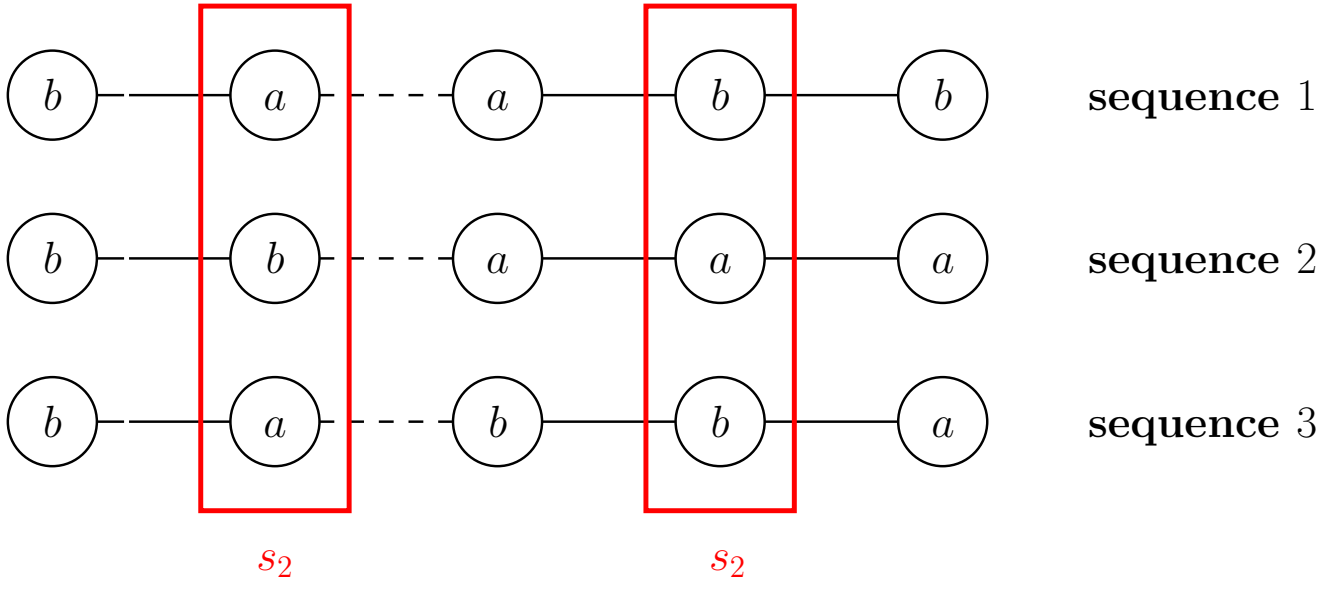
Site patterns

Four site patterns can occur on τ.



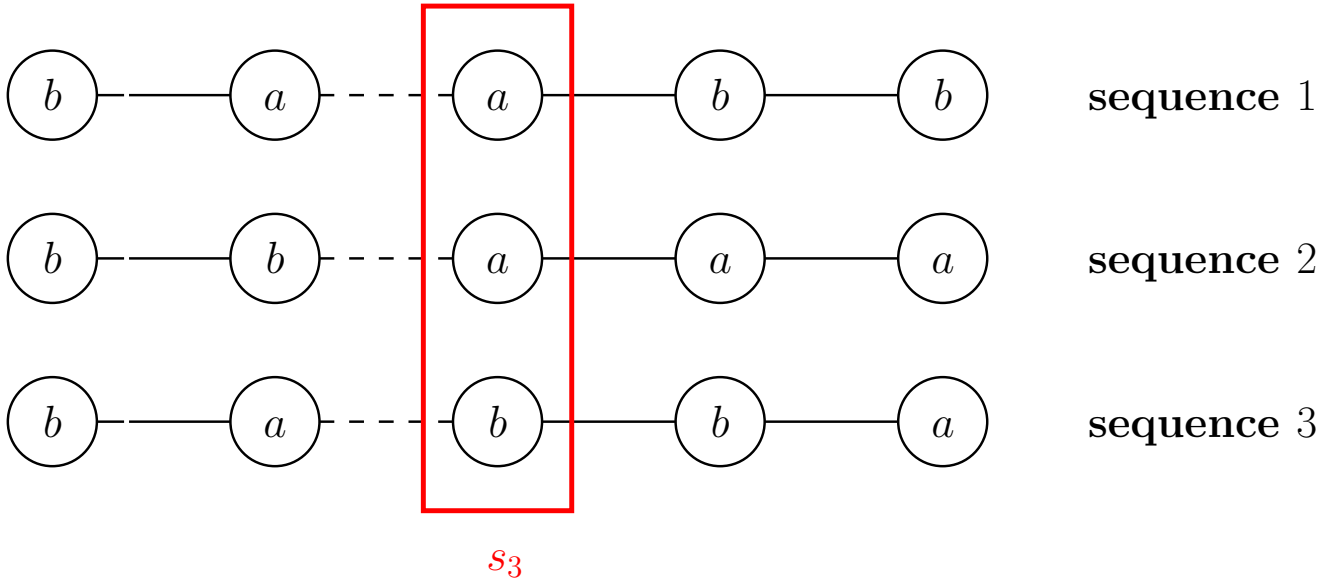
Site patterns

Four site patterns can occur on τ.



Site patterns

Four site patterns can occur on τ.



Computation of $\mathbb{P}(\text{Data}|\text{Tree})$

Let $n_{0:3} = (n_0, n_1, n_2, n_3)$ denote the counts of the different types of site patterns.

Let $p_i(t_e, t_i)$ denote the probability that the site pattern s_i appears on tree R_1 under the 2-state symmetric Markov process, with branch lengths (t_e, t_i).

$$
\begin{aligned}
4p_0(t_e, t_i) &= 1 + \mathrm{e}^{-4t_e} + 2\mathrm{e}^{-4(t_i+t_e)}, \\
4p_1(t_e, t_i) &= 1 + \mathrm{e}^{-4t_e} - 2\mathrm{e}^{-4(t_i+t_e)}, \\
4p_2(t_e, t_i) &= 4p_3(t_e, t_i) = 1 - \mathrm{e}^{-4t_e}.
\end{aligned}
$$

Computation of $\mathbb{P}(\text{Data}|\text{Tree})$

Let $\mathfrak{T} = (T_e, T_i)$ denote a couple of non negative random variables and $P_i = p_i(\mathfrak{T})$.

Likelihood

$$\mathbb{P}(N = n_{0:3}|R_1) = \frac{n!}{n_0!n_1!n_2!n_3!}\mathbb{E}(P_0^{n_0}P_1^{n_1}P_2^{n_2}P_3^{n_3}).$$

Posterior probability distribution on (R_1, R_2, R_3)

Posterior probabilities

$$\frac{\mathbb{P}(R_1|N = n_{0:3})}{\mathbb{P}(R_2|N = n_{0:3})} = \frac{\mathbb{E}(P_0^{n_0} P_1^{n_1} P_2^{n_2+n_3})}{\mathbb{E}(P_0^{n_0} P_1^{n_2} P_2^{n_1+n_3})},$$

and

$$\frac{\mathbb{P}(R_1|N = n_{0:3})}{\mathbb{P}(R_3|N = n_{0:3})} = \frac{\mathbb{E}(P_0^{n_0} P_1^{n_1} P_2^{n_2+n_3})}{\mathbb{E}(P_0^{n_0} P_1^{n_3} P_2^{n_1+n_2})}.$$

Definition

For every $\varepsilon > 0$, *let* $\mathscr{N}_1^{\varepsilon}$ *denote the set of* $n_{0:3}$ *such that, for* $j \in \{2, 3\}$,

$$\mathbb{E}(P_0^{n_0} P_1^{n_1} P_2^{n-n_0-n_1}) \geqslant (2/\varepsilon)\, \mathbb{E}(P_0^{n_0} P_1^{n_j} P_2^{n-n_0-n_j}).$$

For every $n_{0:3} \in \mathscr{N}_1^{\varepsilon}$,

$$\mathbb{P}(R_1 | N = n_{0:3}) \geqslant 1 - \varepsilon.$$

About the parameter r in JC + CpG

At the moment, a prior knowledge of the parameter r is needed to apply the method.

A different approach to estimate all the parameters at the same time, based on maximum likelihood principle is currently developed by Bérard and Guéguen. But the topology of the tree has to be fixed before applying the method.

Insertion/Deletion mechanisms

Substitutions are not the only way to alter DNA sequences. For example, **insertions** add one or several extra nucleotides to the DNA sequence, and **deletions** remove one or several nucleotides from the DNA sequences. One could study stochastic models taking into account these four mechanisms: independent evolutions of the sites, influence of the neighbourhood, insertions and deletions.

Bayesian approach

An open question is to provide necessary conditions for the star paradox to occur. In other words, what priors would prevent the star paradox to occur?

Another line of research is to extend Susko's results on posterior probabilities to non continuous priors, that is, to compute the limit of posterior probabilities when the priors are more general.

Thank you for your attention.

Je vous remercie de votre attention.