Using Stochastic Mapping for model estimation

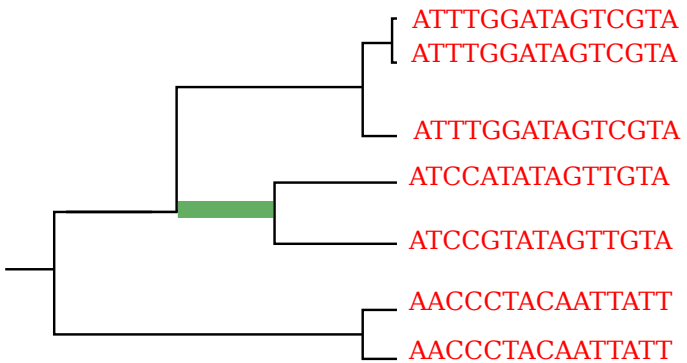
Laurent Guéguen
Lab. Biométrie et Biologie Évolutive - UMR CNRS 5558 – Lyon 1

June 16th 2014

What is mapping ?

Reconstruction of branch history

- Number of given events :
  - Transitions/transversions
  - Towards GC /towards AT
  - Non-synonymous/synonymous
- Time spent in a given state :
  - Stability of amino-acids

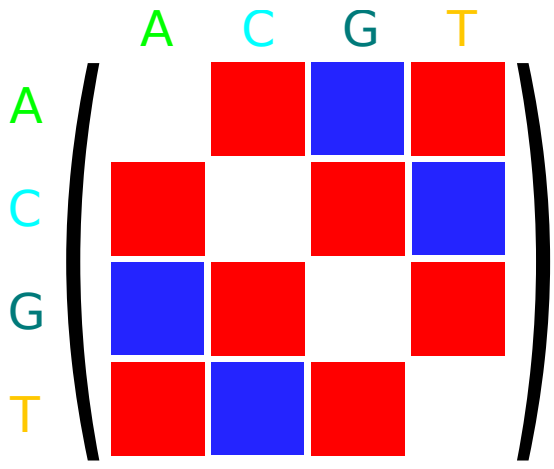


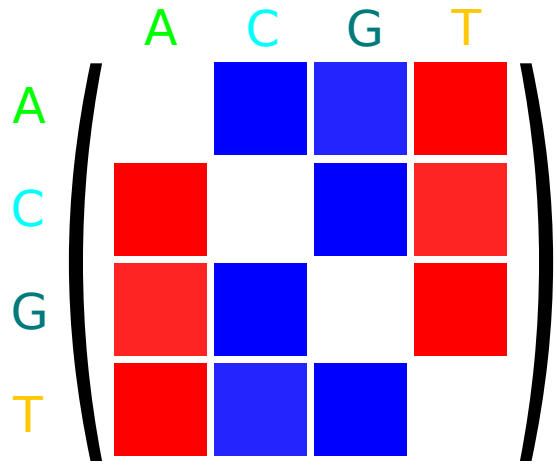
Complementary to ancestral sequence reconstruction (node)

[Nielsen, 2002]

Branch history

An alphabet $\mathcal{A}$ and sets of events $\mathcal{L}_1, \mathcal{L}_2 \subset \mathcal{A} \times \mathcal{A}$.

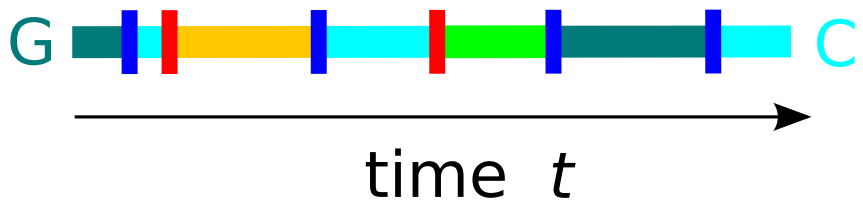




Transitions vs Transversions

Towards GC vs Towards AT

On a branch on a given site σ there is an history :



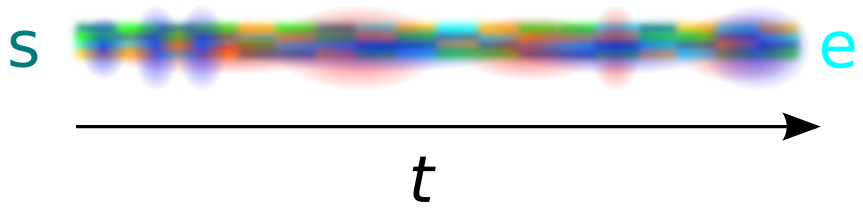
which is **unknown**.

Conditional mapping

In the context of probabilistic modelling :

- Markovian model : $\mathcal{M}$

we infer a distribution of the histories on a branch :



Formula to compute, on a branch of length t, given states s and e :

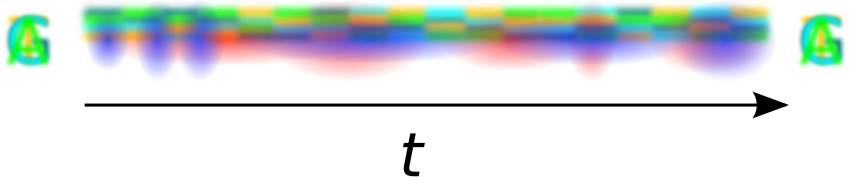
$E(N_{\mathcal{L}}|s, e, t, \mathcal{M})$ expected number of $\mathcal{L}-$events ;

$E(t_a|s, e, t, \mathcal{M})$ expected time spent in state a.

[Minin and Suchard, 2008, Tataru and Hobolth, 2011]

Stochastic mapping

But ancestral sequences are unknown, only the extant sequences $\mathbb{D}$.



We need to integrate on all the ancestral states.
Given a markovian process $\mathcal{P}$, on a branch β, on a site σ, a posteriori joint-probabilities of states :

$$P_\beta(s, e|\mathbb{D}_\sigma, \mathcal{P})$$

$\Longrightarrow$

Expected number of $\mathcal{L}$-events :

$$E_\beta(N_\mathcal{L}|\mathcal{M}, \mathcal{P}, \mathbb{D}) = \sum_{\text{site } \sigma} \sum_{s,e} E(N_\mathcal{L}|s, e, \mathcal{M}) P_\beta(s, e|\mathbb{D}_\sigma, \mathcal{P})$$

Expected time spent in state a :

$$E_\beta(t_a|\mathcal{M}, \mathcal{P}, \mathbb{D}) = \sum_{\text{site } \sigma} \sum_{s,e} E(t_a|s, e, \mathcal{M}) P_\beta(s, e|\mathbb{D}_\sigma, \mathcal{P})$$

Same complexity as likelihood computation $\Longrightarrow$ Fast

Usually $\mathcal{M}$ is the model of $\mathcal{P}$ on branch β.

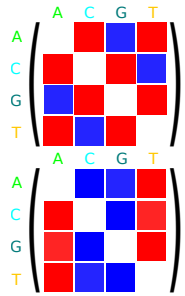
Parameters and Events

Estimation of model parameters from counts

Usual nucleotidic models :

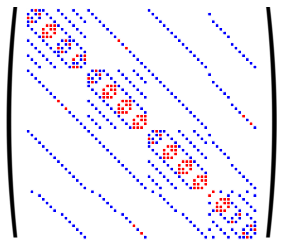
$$\kappa = \frac{\text{Rate of transitions}}{\text{Rate of transversions}}$$

$$\theta = \frac{\text{Rate of substitutions towards GC}}{\text{Rate of substitutions towards GC} + \text{Rate of substitutions towards AT}}$$



Usual codon model (Yang & Nielsen 1998) :

$$\omega = \frac{dN}{dS}$$



But : Rates are required, ie counts **"per relevant site"**.

$\Rightarrow$ take into account ancestral sequences and models ;
$\Rightarrow$ More relevant information on the past substitution process.

dN and dS

Usual measures of evolution of codon sequences :

dN number of non-synonymous substitutions **per non-synonymous site**.

dS number of synonymous substitutions **per synonymous site**.

What does it mean ? It depends on :

- the ancestral sequence
- the model

[Goldman and Yang, 1994] : Number of (non-)synonymous substitutions that would be performed by a neutral model

$\Longrightarrow$ Normalization by the rates of a **neutral** model $\omega = 1$.

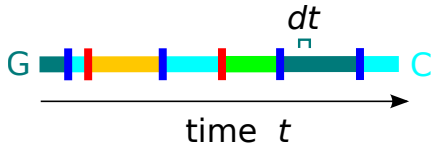
Model ability

The expected amount of substitutions in $\mathcal{L}$ that a model is able to perform on a branch.

Generator of model $\mathcal{M}$: $\mathcal{Q}$

Instantaneous model ability :

During a small time dt, with letter $X(\tau)$ at time τ

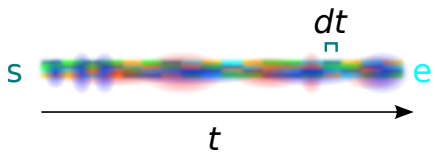


Sum of the substitution rates of events in $\mathcal{L}$:

$$\sum_{b \in \mathcal{A}; (X(\tau), b) \in \mathcal{L}} \mathcal{Q}_{X(\tau),b} dt = \mathcal{Q}_{X(\tau),\mathcal{L}} dt$$

with $\mathcal{Q}_{a,\mathcal{L}} = \sum_{b \in \mathcal{A}; (a,b) \in \mathcal{L}} \mathcal{Q}_{a,b}$.

Since the history is not known, given states s and e :



$$\sum_{a \in \mathcal{A}} \mathcal{Q}_{a,\mathcal{L}} P(X(\tau) = a | s, e, t, \mathcal{M}) dt$$

Model ability

So, the ability of $\mathcal{M}$ along the branch is :

$$\begin{aligned} A^{\mathcal{L}}(\mathcal{M}|s,e,t) &= \int_{\tau=0}^{t} \sum_{a\in\mathcal{A}} \mathcal{Q}_{a,\mathcal{L}} P(X(\tau)=a|s,e,t,\mathcal{M})dt \\ &= \sum_{a\in\mathcal{A}} \mathcal{Q}_{a,\mathcal{L}} \int_{\tau=0}^{t} P(X(\tau)=a|s,e,t,\mathcal{M})dt \\ &= \sum_{a\in\mathcal{A}} \mathcal{Q}_{a,\mathcal{L}} E(t_a|s,e,t,\mathcal{M}) \end{aligned}$$

And, given a markovian process $\mathcal{P}$, on a branch β,
a posteriori model ability in $\mathcal{L}$:

$$A^{\mathcal{L}}_{\beta}(\mathcal{M}|\mathbb{D},\mathcal{P}) = \sum_{\text{site }\sigma} \sum_{s,e} \sum_{a\in\mathcal{A}} Q_{a,\mathcal{L}} E(t_a|s,e,\mathcal{M}) P_{\beta}(s,e|\mathbb{D}_{\sigma},\mathcal{P})$$

Same complexity as likelihood computation $\Longrightarrow$ Fast

Parameter estimation

1. Perform substitution mapping ;
2. Normalize of the counts :
   - Use the same model as the one defined in process $\mathcal{P}$;
   - Change the parameter value by its "null" value : $\mathcal{M}_0$;
   - Divide the count by the ability of the "null" model :

$$\frac{E_\beta(N_\mathcal{L}|\mathcal{M},\mathcal{P},\mathbb{D})}{A^\mathcal{L}_\beta(\mathcal{M}_0|\mathbb{D},\mathcal{P})}$$

3. Estimate the parameters from the normalized counts.

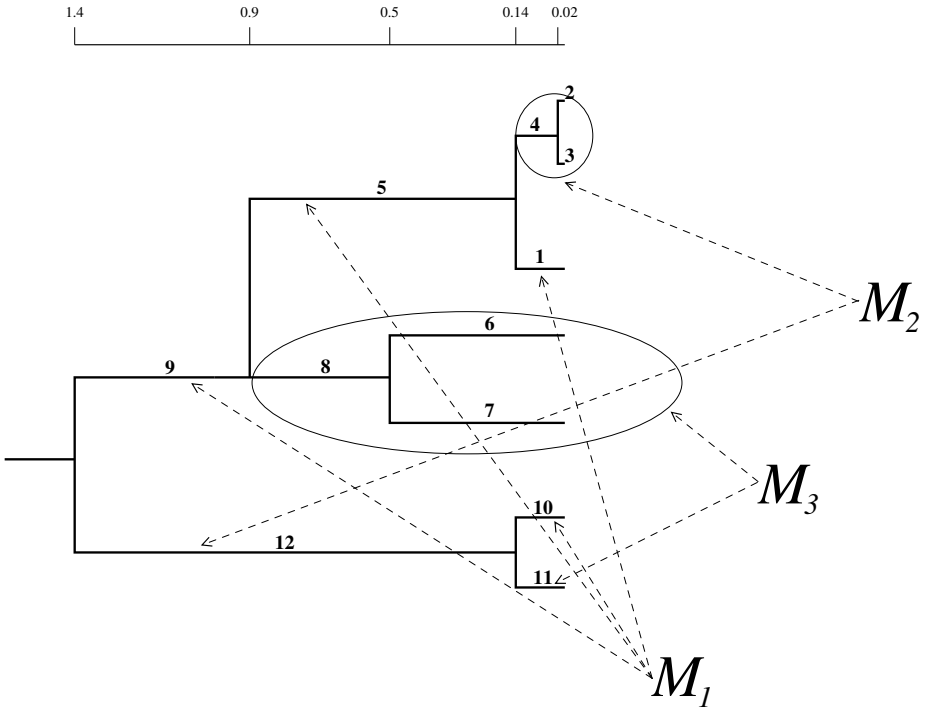
For example, on YN98(F1X4) :

parameter	null value	$\mathcal{L}_1$ – events	$\mathcal{L}_2$ – events	formula
κ	1	transitions	transversions	$\frac{\mathcal{L}_1}{\mathcal{L}_2}$
θ	0.5	towards GC	towards AT	$\frac{\mathcal{L}_1}{\mathcal{L}_1+\mathcal{L}_2}$
ω	1	nonsynonymous	synonymous	$\frac{\mathcal{L}_1}{\mathcal{L}_2}$

Simulations

100 codon alignments with YN98(F1X4) modelling

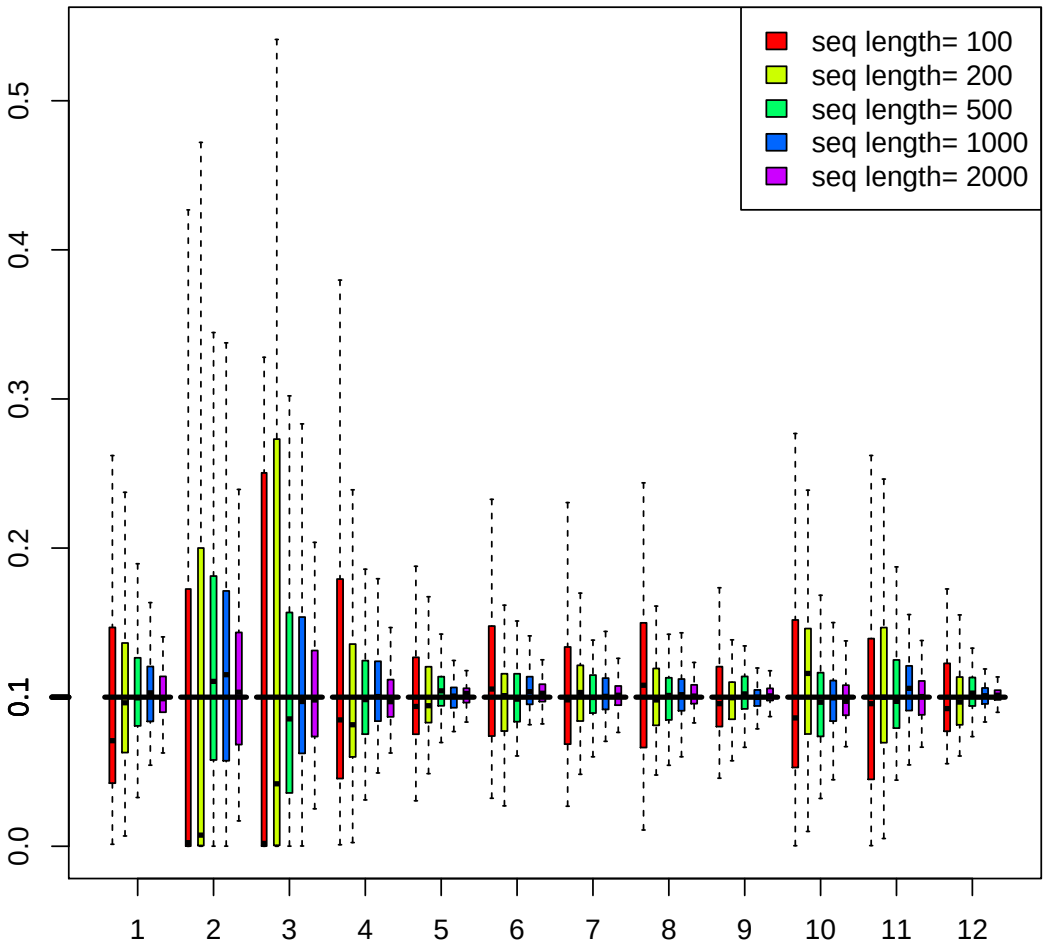
1. Simulation of an alignment with **three** models $\Rightarrow \mathbb{D}$;

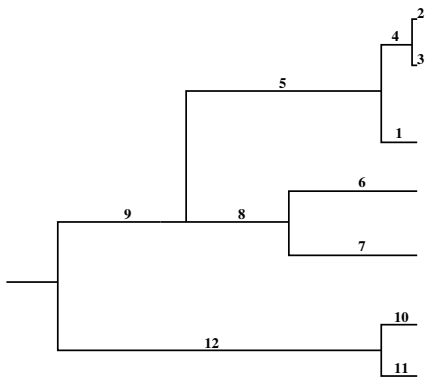


2. Maximum likelihood inference using **one** model $\Rightarrow \mathcal{P}$;
3. Estimation of the parameters of the **three** models.

$\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}_3$

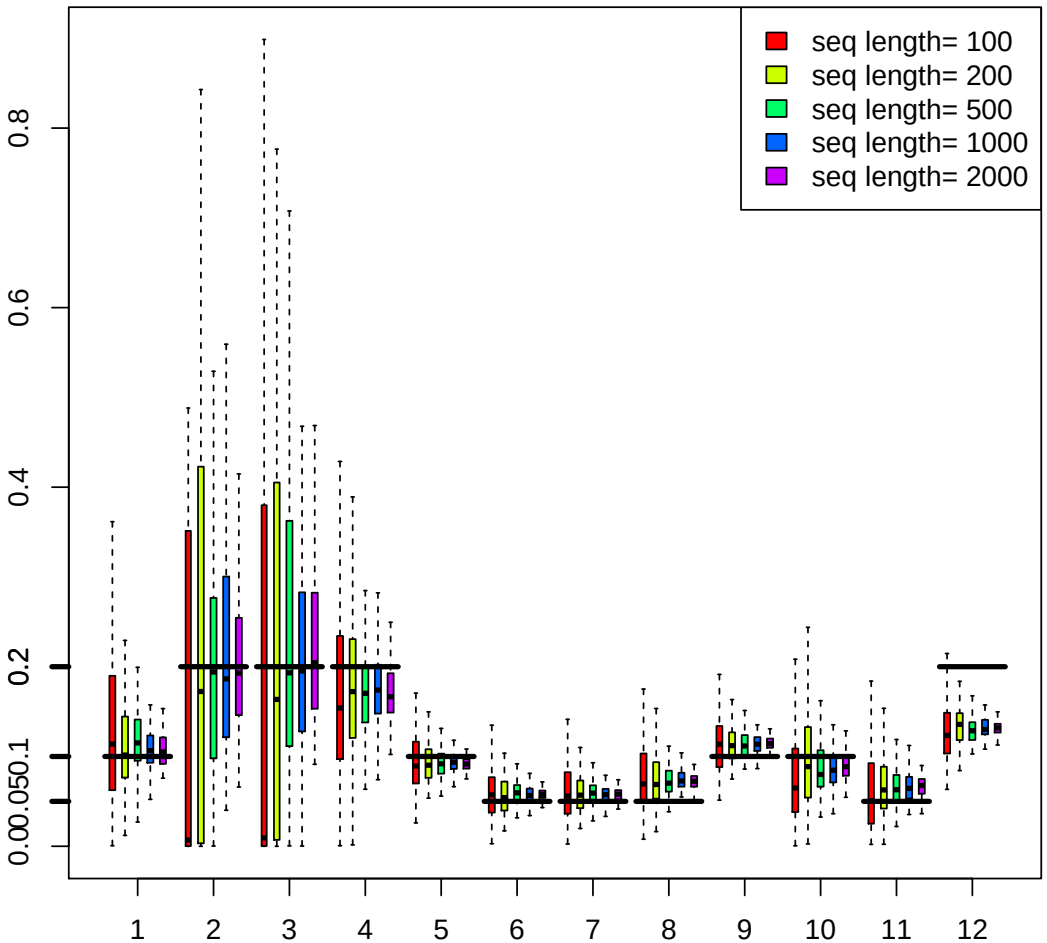
Estimation of ω

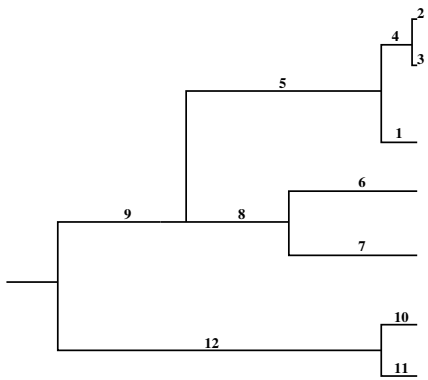




$\mathcal{M}_1 \neq \mathcal{M}_2 \neq \mathcal{M}_3$

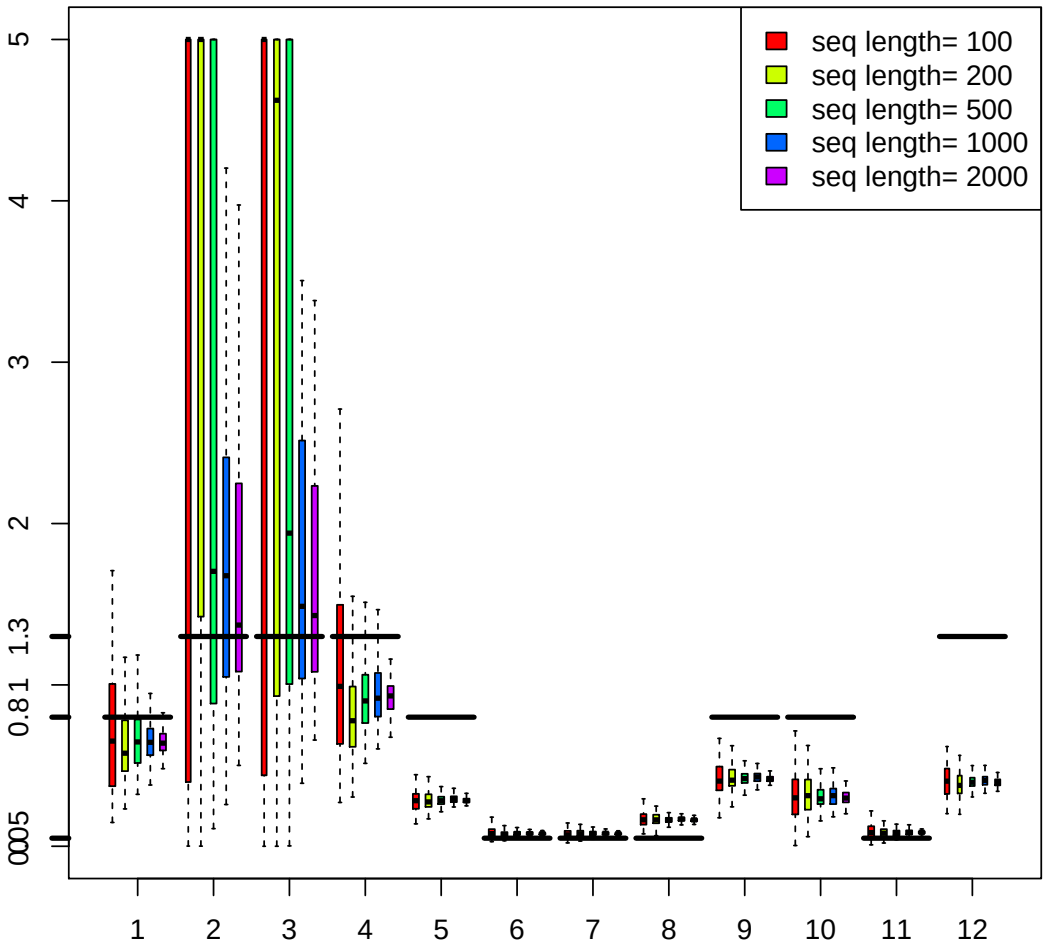
Estimation of ω

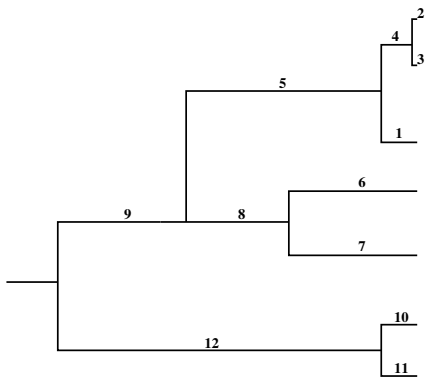




$\mathcal{M}_1 \neq \mathcal{M}_2 \neq \mathcal{M}_3$

Estimation of ω





To conclude

- Normalization of mapping counts to get better information of the substitution process, taking into account :
  - a posteriori ancestral sequences
  - bias in the model
- $\Rightarrow$ New formula for dN and dS.
- Unbiased estimator with good model

- On the way to get non-homogeneity