Phylogenetic Reconstruction by Bayesian Analysis of a Non-stationary Model of Biological Sequences Evolution

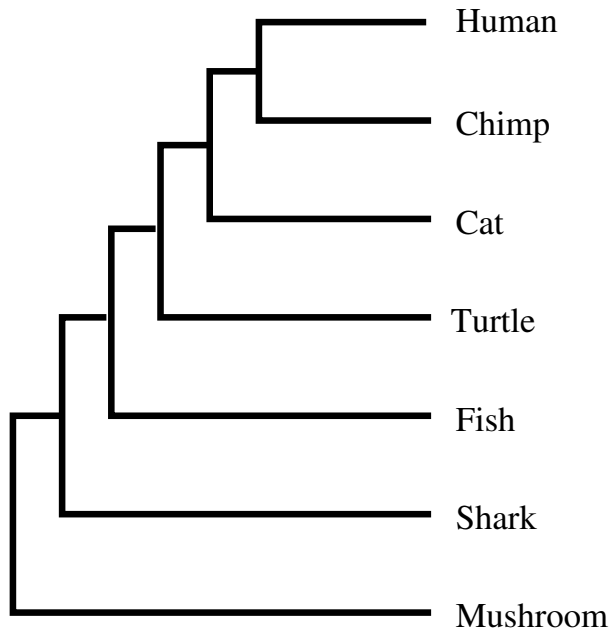
Samuel Blanquart

Methods and Algorithms for Bioinformatic Project
LIRMM CNRS

January 22, 2006

Contents

Phylogenetic Trees and Sequences Alignment

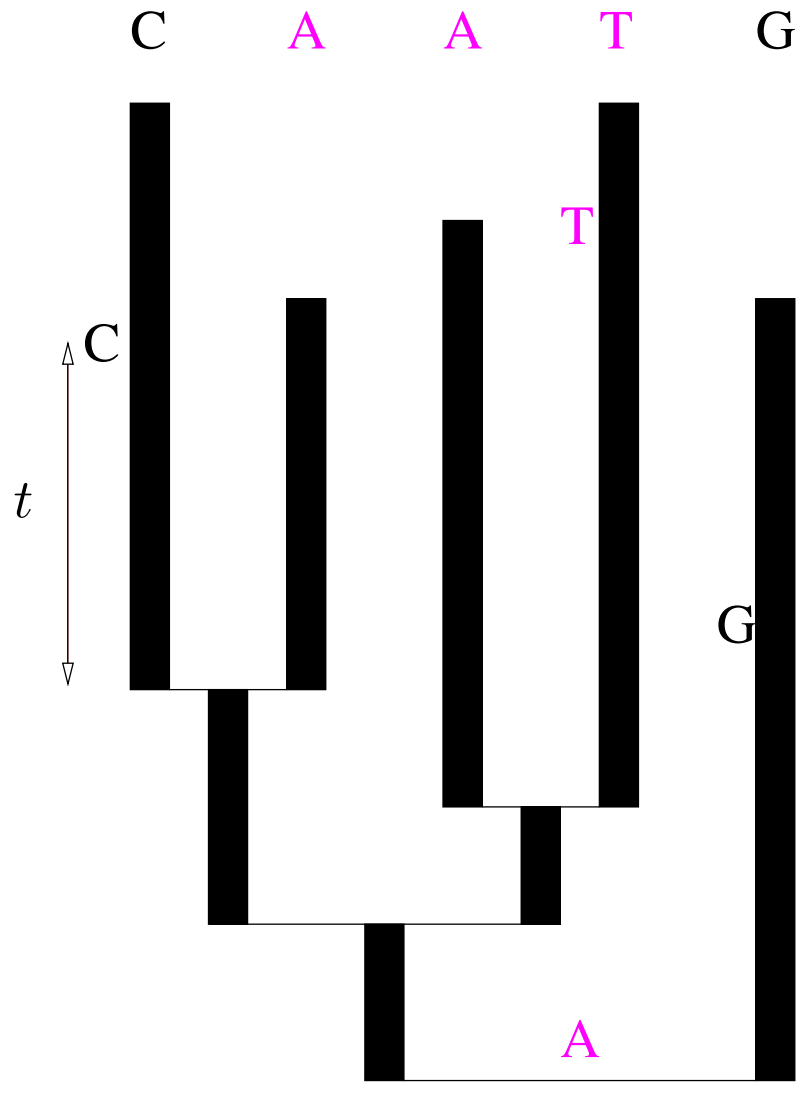


human	T GGGG C GAAGTCGTAACAAGGTA G C C
chimp	T GGGG T GAAGTCGTAACAAGGTA A C T
cat	G GGGG C GAAGTCGTAACAAGGTA G C C
turtle	T GGGG T GAAGTCGTAACAAGGTA G C C
fish	T GGGG C GAAGTCGTAACAAGGTA G C T

Stationary Models

$$P(A \rightarrow C) = [e^{tQ}]_{AC}$$

Profile 



Example: MrBayes [Huelsenbeck and Ronquist, 2001]

Extent Non-stationary Models

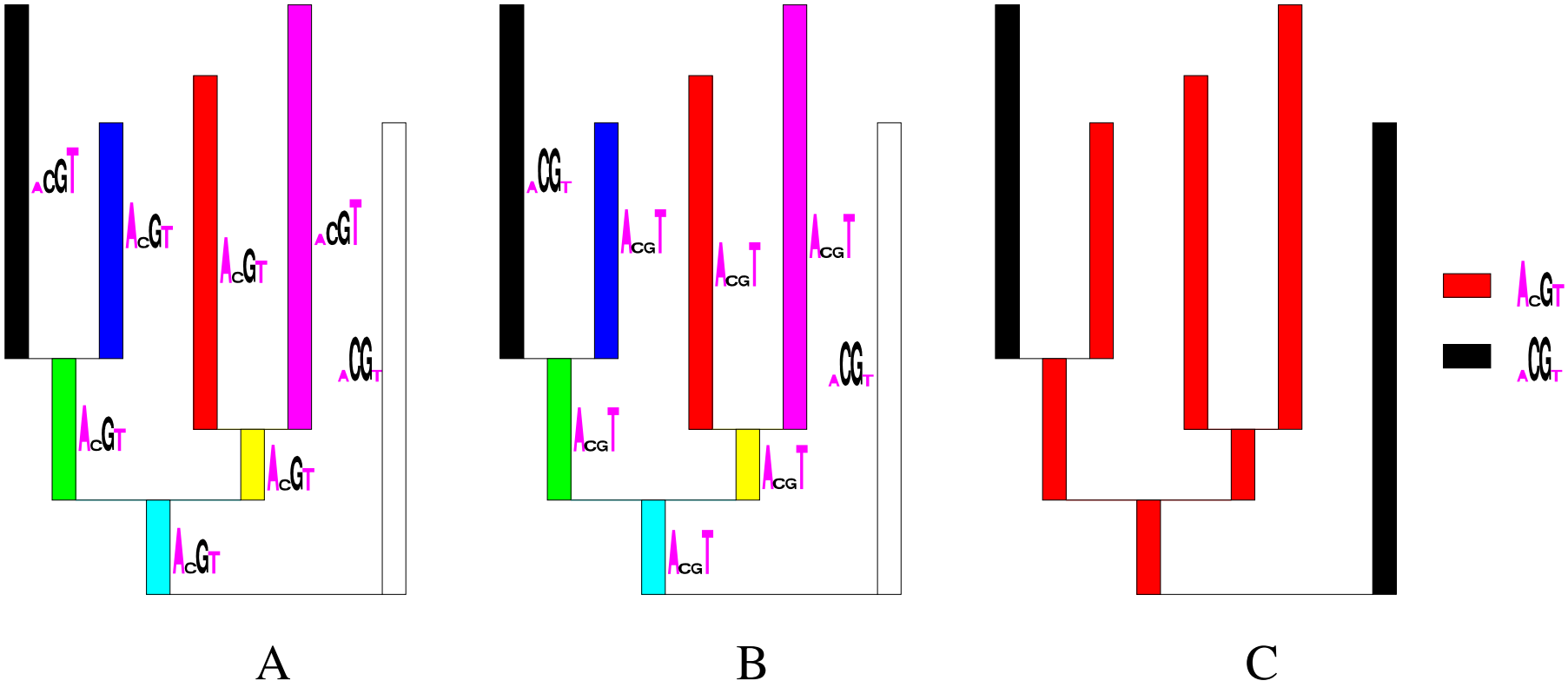
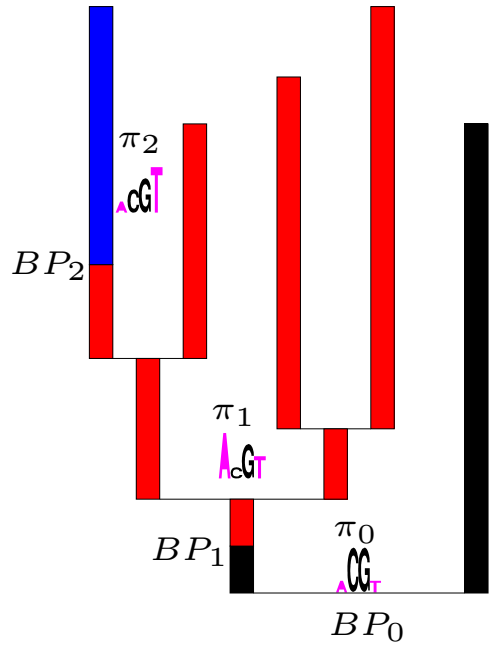


Figure: figure 1 A: [Yang and Roberts, 1995],
figure 1 B: [Galtier and Gouy, 1998],
figure 1 C: [Foster, 2004].

Non-stationary Model Description

The profiles are punctually modified along the tree, following a stochastic Poisson process of rate ϵ, at points which we call breakpoints, or BP:



The vector of the model parameters, θ, is defined by:

$$\theta = \{\tau, N, \pi, X, \epsilon\} \tag{1}$$

Priors of the Model Parameters

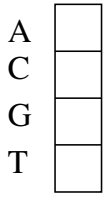
We demonstrate that a particular distribution of N breakpoints, on a given topology, has the following probability density:

$$P(N, \mathbf{X}) = \left(\frac{e^{-\epsilon T} (\epsilon T)^N}{N!} \right) \left(\frac{N!}{T^N} \prod_{j=1}^{2J-3} t_j^{n_j} \right), \qquad (2)$$

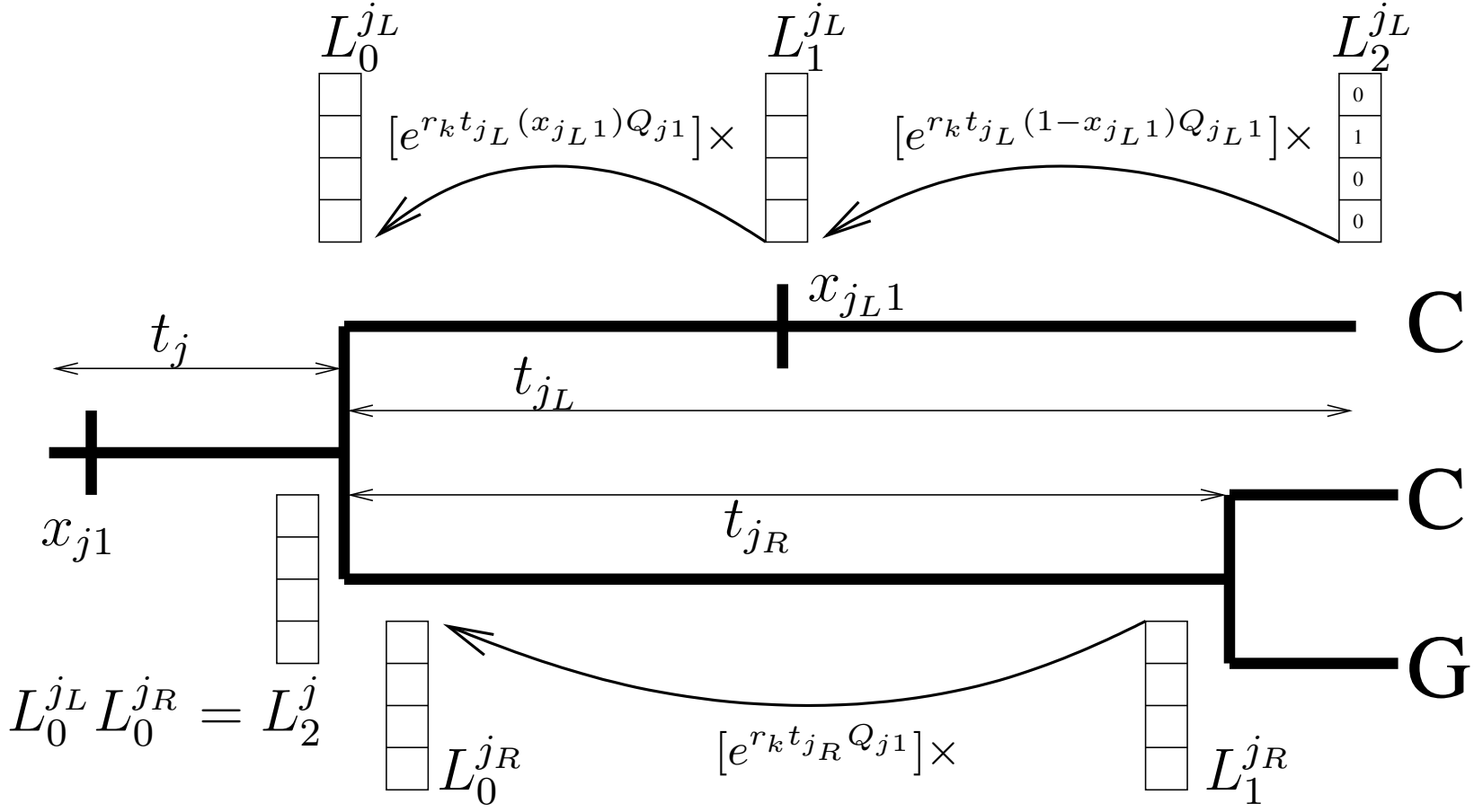
where T is the total tree length. We also define probability distributions for all other parameters, topology (Uniform), rate across sites (Gamma), profile shapes (Uniform), hyperparameter ϵ (exponential), ...

Likelihood Computation

We adapt Felsenstein's pruning algorithm ([Felsenstein, 1981]) to compute recursively the likelihood (L) between breakpoints:



Partial likelihood vector



Bayes Theorem

$$P(\theta \mid D, M) = \frac{P(D \mid \theta, M) \times P(\theta \mid M)}{P(D \mid M)}, \tag{3}$$

where

- M is a model,
- θ is the vector parameters of the model M,
- D are the data,
- $P(\theta \mid D, M)$ is the posterior,
- $P(D \mid \theta, M)$ is the likelihood,
- $P(\theta \mid M)$ is the prior probability,
- $P(D \mid M)$ is the Bayes factor.

MCMC Sampling General Principle

One can sample posterior probability density, induced by the data on the model parameters, with a Markov Chain Monte Carlo. We use the standard Metropolis-Hasting algorithm:

1. Modify θ into θ^*, according to the MCMC kernel q,
2. Evaluate the Metropolis-Hasting ratio:
   $MH = \frac{P(\theta^*|D)}{P(\theta|D)} \times \frac{q(\theta^*,\theta)}{q(\theta,\theta^*)}$,
3. Accept or refuse θ^* with probability
   $P_{accept}(\theta^*) = min(1, MH))$,
4. If θ^* is accepted, $\theta = \theta^*$,
5. go to step 1

MCMC Stochastic Kernel q

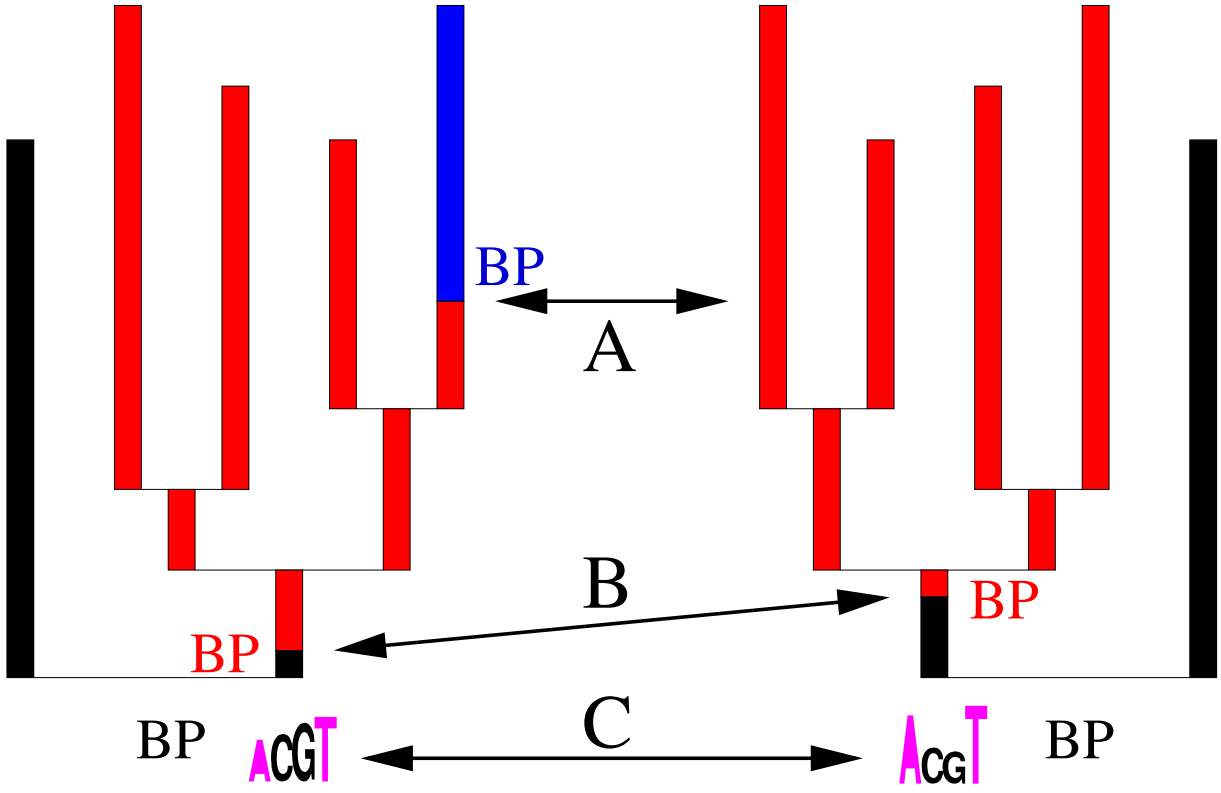
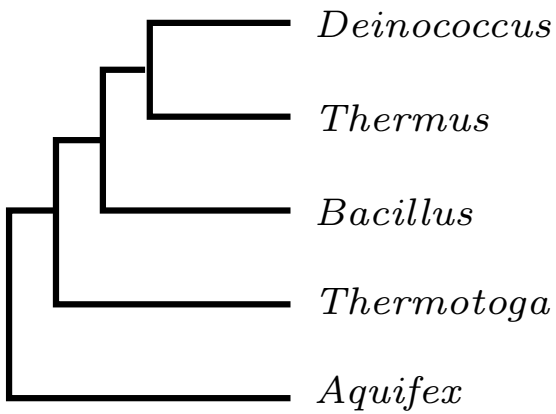


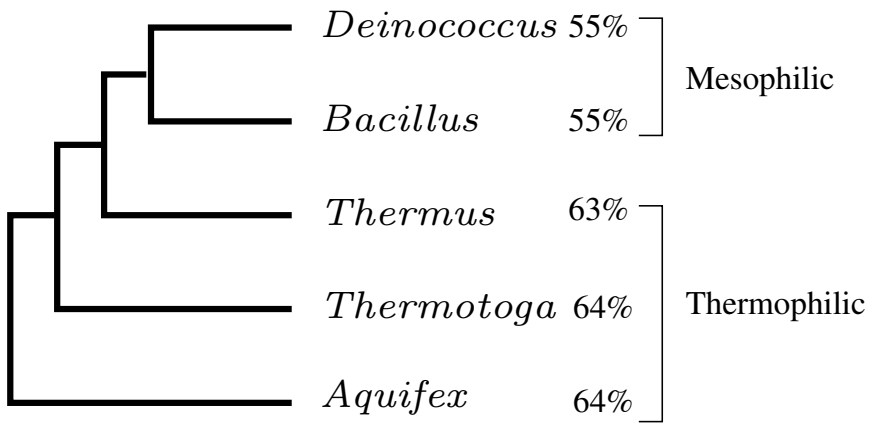
Figure: Methods used to update the breakpoint structure, figure 2 A: create or delete a breakpoint, figure 2 B: update a breakpoint position, figure 2 C: update a breakpoint profile.

Material

For 5 bacterial 16S rRNAs, two candidate topologies are available, the first (τ_1) is known with confidence to be true ([Murray, 1991], [Gupta, 1998], [Eisen, 1995], [Galtier and Gouy, 1998], [Olsen et al., 1994], [Foster, 2004]), whereas the other (τ_2) is an artifact:



τ_1



τ_2

Results, Model Parameters Posterior Values

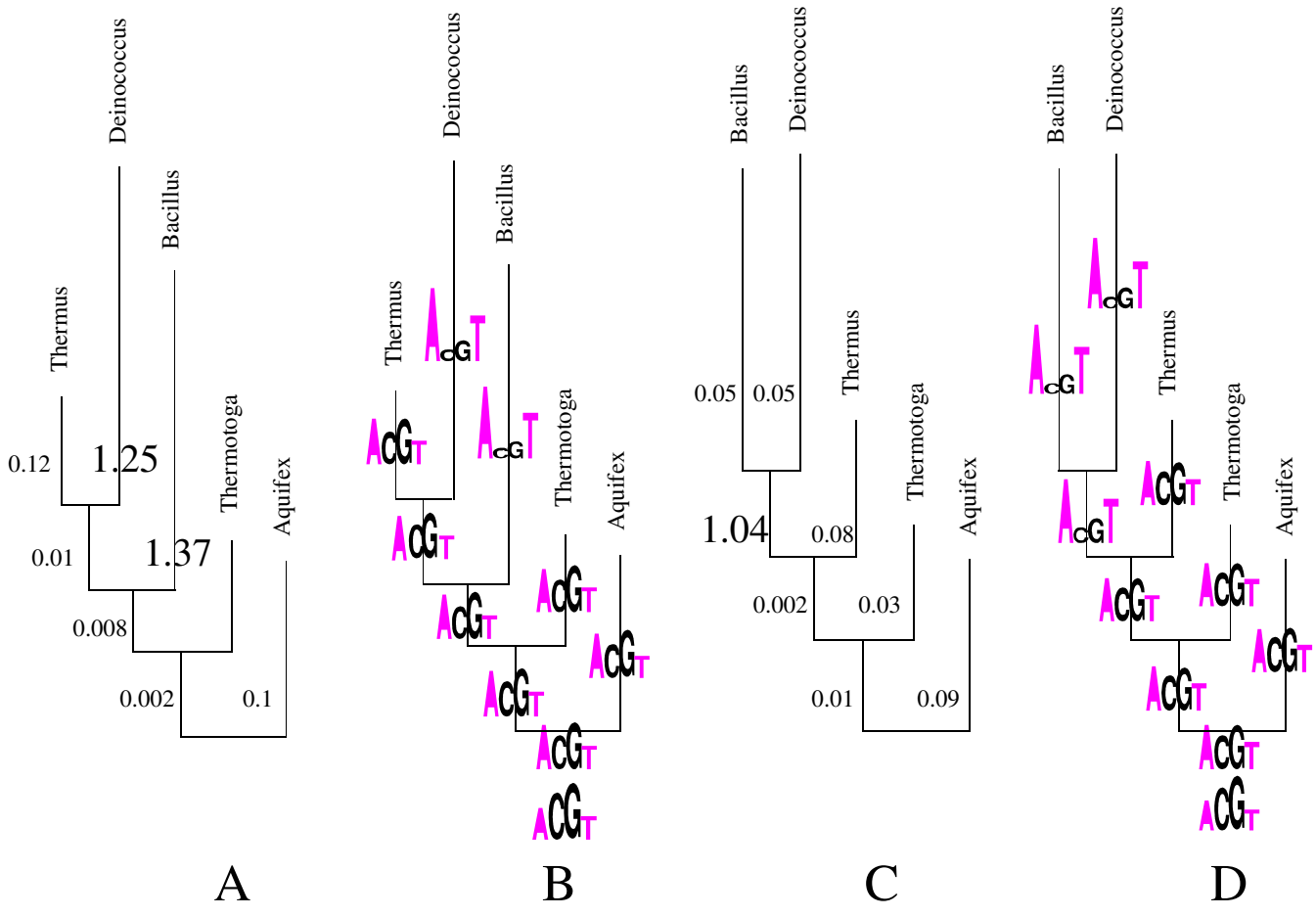


Figure: Mean, over all MCMC sample, posterior number of breakpoints (figure 3 A, 3 C) and mean posterior profiles (figure 3 B, 3 D), for the fixed true (figure 3 A, 3 B) and artifact (figure 3 C, 3 D) topologies.

Results, Comparing Fit of Models

We compare the fit of non-stationary models BP and YR ([Yang and Roberts, 1995]), and of our model stationary configuration $STAT$, versus the stationary model, by thermodynamic integration ([Gelman et al., 2004], [Ogata, 1989]):

	lower bound	upper bound
BP	63.24	63.27
YR	58.01	59.04
$STAT$	0.20	0.39

Positives values of the Bayes factor show that the stationary model is rejected.

Results, the Over-parametrization Issue

Model comparison, by AIC score evaluation ([Akaike, 1974]): $AIC =< log(L) > -\frac{1}{2}k$, where k is the number of free parameters.

	15 species	30 species
BP	7742 (5 BP)	12222 (7 BP)
YR	7768 (28 BP)	12269 (58 BP)
YR-BP	26	47

Conclusion

The stationarity assumption does not apply when analysing phylogenetic relationships between strongly biased sequences. We show that it is then better to use a non-stationary model.

We are currently studying, using our non-stationary model, some other datasets, for which we obtain more accurate results than with standard models.

Bibliography


Akaike, H. (1974).
A New Look at the Statistical Model Identification.
IEEE Transactions on Automatic Control, 19(6):716–722.


Eisen, J. A. (1995).
The RecA Protein as a Model Molecule for Molecular Systematic Studies of Bacteria: Comparison of Trees of RecAs and 16S rRNAs from the same Species.
Molecular Biology and Evolution, 41(6):1105–1123.


Felsenstein, J. (1981).
Evolutionary Trees from DNA Sequences: A Maximun Likelihood Approach.
Molecular Evolution, 17(6):368–376.


Foster, P. G. (2004).
Modeling Compositional Heterogeneity.
Systematic Biologists, 53(3):485–495.


Galtier, N. and Gouy, M. (1998).