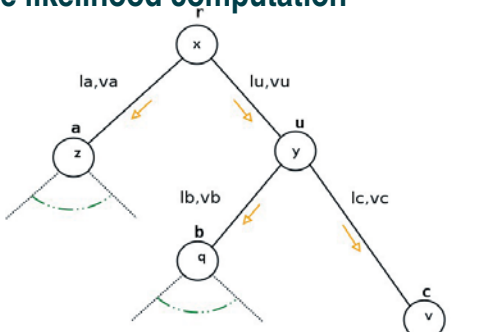
Efficiently finding the most likely tree under non-reversible models of evolution

Bastien Boussau, Manolo Gouy

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1, 43 bd 11 nov, 69622, Villeubanne Cedex, France.

Molecular phylogeny estimation is best fulfilled in a **statistical framework**, with **Bayesian** or **Maximum Likelihood** methods. Modeling sequence evolution properly is of the utmost importance, for infering past events under inappropriate hypothesis can lead to erroneous conclusions. It is commonly considered that the process of evolution is **homogeneous and stationary**, which means that sequence composition is supposed to be constant over the whole phylogeny. The mere analysis of nowadays sequences clearly shows that such a simplification is **highly unrealistic**, but the algorithmic complexities yielded by its dismisal have discouraged people from using non-homogeneous, non-stationary models of evolution (NHNS models). More precisely, while most models of evolution are **reversible**, meaning that no evolutionary direction and no root are constrained, these **NHNS models are non-reversible**, and make it necessary to define a root *a priori* and to keep it throughout the whole search for likely trees. **Here, we explain that it is possible with only minor modifications to use cutting-edge algorithms with non-reversible models of evolution, and adapt the NHNS model of Galtier and Gouy (1998) to the algorithm of phyml (Guindon and Gascuel, 2003).**

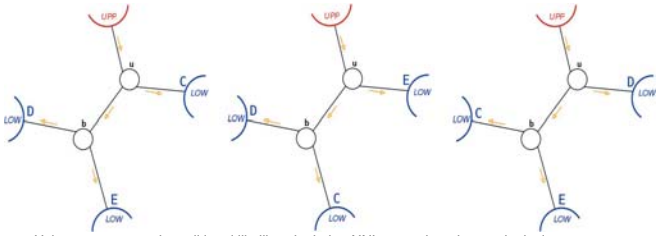
I) Tree likelihood computation



$$
\begin{aligned}
L_s &= \sum_{x\in\Omega} P(r=x) \\
&\times \sum_{y\in\Omega} P_{xy}(l_u, v_u) \\
&\times \sum_{z\in\Omega} P_{xz}(l_a, v_a) L_{s,low(ra)}(a=z) \\
&\times \sum_{q\in\Omega} P_{yq}(l_b, v_b) L_{s,low(ub)}(b=q) \\
&\times \sum_{v\in\Omega} P_{yv}(l_c, v_c) L_{s,low(uc)}(c=v)
\end{aligned}
$$

II) Efficient likelihood computation

When exploring the topological space, heuristics need to estimate the likelihoods of many different topologies. In order to save time, any unnecessary computation is avoided.
In this respect, conditional likelihoods are defined that contain the likelihoods of subtrees.



Using pre-computed conditional likelihoods during NNIs to explore the topological space

These conditional likelihoods are defined as follows.

- For a leaf *c*:

$$
L_{s,low(uc)}(c=v) = \begin{cases} 1 & if\ base\ v\ is\ at\ site\ s\ of\ leave\ c \\ 0 & otherwise \end{cases}
$$

- For a subtree whose root is in *u*:

$$
L_{s,low(ru)}(u=y) = \sum_{q\in\Omega} P_{yq}(l_b, v_b) L_{s,low(ub)}(b=q) \times \sum_{v\in\Omega} P_{yv}(l_c, v_c) L_{s,low(uc)}(c=v)
$$

With a non-reversible model of evolution, some subtrees contain the fixed root, for which the above formulas do not apply.
Then the optimization used in most efficient algorithms cannot be applied to non-reversible models of evolution, which have thus been excluded from most phylogeny packages.

III) Efficient likelihood computation in the non-reversible case

When dealing with non-reversible models of evolution, it is necessary to account for the position of the root: root base distribution is then included in upper conditional likelihoods.

- The upper likelihood corresponding to branch *ru* is as follows:

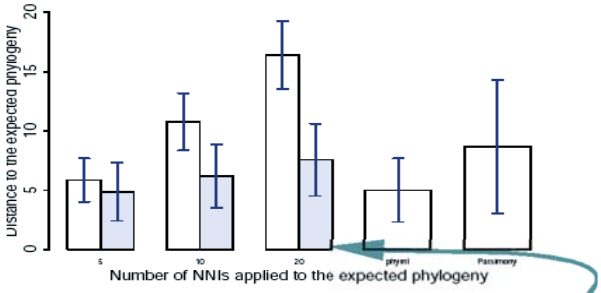
$$
L_{s,upp(ru)}(r=x) = P(r=x) \times \sum_{z\in\Omega} P_{xz}(l_a, v_a) L_{s,low(ra)}(a=z)
$$

- Underlying branches' upper conditional likelihoods can be defined recursively:

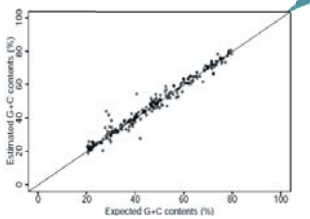
$$
L_{s,upp(ub)}(u=y) = \sum_{x\in\Omega} P_{xy}(l_u, v_u) L_{s,upp(ru)}(r=x) \times \sum_{v\in\Omega} P_{yv}(l_c, v_c) L_{s,low(uc)}(c=v)
$$

IV) nhPhyml, phyml adapted to Galtier and Gouy's model (1998)

a) Ability to explore the topological space



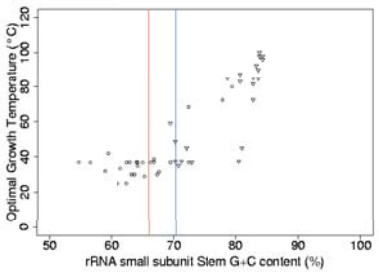
b) Ability to estimate the ancestral G+C content



→ nhPhyml is able to explore the topological space and to estimate the ancestral G+C content.

V) The Last Universal Common Ancestor rRNA G+C content, and the question of its optimal growth temperature

rRNA stem G+C content is known to be correlated to the host optimal growth temperature. Estimating the ancestral rRNA stem G+C content could hence give a hint about the ancestral optimal growth temperature.
We inferred the ancestral SSU and LSU stem G+C contents, placing the root on the branch leading to Archaea, Bacteria or Eukaryotes. The two significantly most likely rootings place **LUCA among mesophilic or thermophilic species.**



Conclusion and Perspectives

- We have reported that **efficient maximum likelihood computation** is not limited to reversible models of evolution but **can be used with non-reversible models of evolution**. This considerably broadens the range of evolutionary models that can be used to search for phylogenies.
- We developped nhPhyml, an adaptation of phyml's architecture to the non-homogeneous, non-stationary model of Galtier and Gouy. Its abilities to explore the topological space and to estimate the ancestral G+C contents have been investigated and suggest that **NHNS models could be useful for phylogeny estimation.** We now plan to adapt nhPhyml to protein-coding sequences through the use of a codon model.
- Galtier and Gouy's model still **does not hint for a hyperthermophilic LUCA.**

Bibliography:
Galtier, N., Gouy, M. (1998). "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneousmodel of DNA sequence evolution for phylogenetic analysis.Mol. Biol. Evol. 15: 871-879.
Galtier, N., Tourasse, N., Gouy, M. (1999). « A non-hyperthermophilic common ancestor to extant life forms ». Science, 283: 220-221
Guindon, S., Gascuel, O. (2003). « A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. ». Syst. Biol., 52: 696-704.