Research in molecular phylogeny aims at reconstructing historical relations between genes or species while trying to capture the true nature of the evolutionary process itself. Both can be estimated at the same time through the use of statistical modeling. Maximum likelihood or the Bayesian framework permit the estimation of parameters of the evolutionary model such as the transition/transversion ratio, the equilibrium base composition, and the tree itself, topology and branch lengths included. Optimizing all these parameters is computationally intensive: the number of possible topologies increases factorially with the number of taxa considered, which makes it necessary to use heuristics when exploring the space of tree topologies. Most recent algorithms (e.g., PhyML [Guindon and Gascuel, 2003], RAxML [Stamatakis et al., 2005]) are able to find trees with excellent likelihood scores for hundreds of sequences, but only with reversible models of evolution. All these reversible models are homogeneous and stationary, i.e., suppose that state evolution is constant all over the tree. If this hypothesis were true, sequences sharing a common ancestor would have the same expected base frequencies.

More precisely, a process of evolution is homogeneous when the state distribution probability simply depends on the time separating it from a given past state distribution probability and not on the branch in the tree: homogeneity is the feature of a process of evolution that is constant in pattern over the whole tree. On the other hand, stationarity is the feature of a process of evolution that keeps the state distribution probability constant over the whole tree: the probability to draw a given state is the same wherever on the tree the sampling is done. A process can be stationary and not be homogeneous, as is the case for Ziheng Yang’s codon model in which the nonsynonymous-to-synonymous ratio varies across branches while codon equilibrium frequencies remain constant all over the tree (Yang, 1998). On the contrary, nonstationarity induces nonhomogeneity, as the process of evolution depends upon the equilibrium frequencies.

The analysis of extant sequences shows that homologous genes vary widely in their composition. As they all stem from a common ancestor, this evidences that sequence evolution is at least not stationary: two sequences in two different species or at two different periods evolve towards different compositions.

The use of nonhomogeneous and nonstationary models that account for this variability in evolution permits minimizing compositional biases and hence improving phylogenetic reconstructions (Galtier and Gouy, 1998; Tarrio et al., 2001; Herbeck et al., 2005). Unfortunately, removing the homogeneity and stationarity hypotheses implies abandoning reversibility, and hence prevents one from using the most efficient algorithms, when those particularly variable-rich models would most eagerly need it.

In this article we show in the general case that it is possible to use recent algorithms with nonreversible models of sequence evolution. We first explain how the likelihood of a tree is computed and how the reversibility property is used in recent heuristics to avoid dispensable calculations during tree space search. We then prove that the same computational trick can be used with nonreversible models of evolution through a reorganization of the way likelihood is computed. As reversible models of evolution can also be used in this framework, this work can be considered as a generalization of the usual formulas, which considerably broadens the amount of models that can be used for a phylogenetic analysis, as all nonhomogeneous as well as nonstationary models can be used.

Eventually we report nhPhyML, the first adaptation of PhyML (Guindon and Gascuel, 2003), a very fast and efficient algorithm, to a nonreversible model of evolution; namely, the implementation of Tamura’s model at each branch of a tree introduced by Galtier and Gouy (Tamura, 1992; Galtier and Gouy, 1998). We also introduce nhPhyML-Discrete, an approximation of nhPhyML. This implementation shows better performance than nhPhyML in the exploration of the space of tree
topologies and similar accuracy in the estimation of the ancestral G+C content. Eventually, we apply nhPhyML to ribosomal RNA slowly evolving sites and conclude that a wider taxonomic sampling than in Galtier et al. (1999) still does not support a hyperthermophilic last universal common ancestor.

**Computing the Likelihood of a Tree from Any Node under a Nonreversible Model of DNA Sequence Evolution**

**Computing the Likelihood of a Tree**

We first explain how one computes the likelihood of a phylogenetic tree with DNA sequences using the following example (Fig. 1).

Most commonly, sites are supposed to evolve independently of each other: a site does not depend on its neighbors' states but only on its past state. As a consequence, the likelihood of a tree for a whole sequence is obtained by multiplying all the likelihoods obtained at single sites.

The likelihood $L_s$ of the tree given in Figure 1 for a single site $s$ is computed as follows:

$$L_s = \sum_{x \in \Omega} \left( P(R = x) \times \sum_{z \in \Omega} \left[ P_{xz}(l_A, v_A)L_{s, \text{low}(RA)}(A = z) \right] \right)$$

where $P_{xy}(l_A, v_A)$ is the probability for base $x$ to change into base $y$ along a branch of length $l_A$ and other evolutionary parameters $v_A$. $P(R = x)$ is the probability to have base $x$ at the root $R$, and $\Omega = \{A, T, C, G\}$ is the set of possible DNA bases. $L_{s, \text{low}(RA)}(A = z)$ is the lower conditional likelihood (Felsenstein, 1981) of observing the data downstream from branch $RA$ conditionally on the underlying subtree and on having base $z$ at node $A$. For each subtree, one can define four conditional likelihoods, one for each DNA base. Once these conditional likelihoods have been computed for a subtree, as long as its topology and branch lengths do not change, they can be re-used if one moves the whole subtree around the topology. This property is used in recent heuristics to search for the most likely phylogenetic tree. These conditional likelihoods are defined as lower, in the sense that they do not contain the root.

Lower conditional likelihoods are defined recursively. For a leaf $C$:

$$L_{s, \text{low}(UC)}(C = v) = \begin{cases} 1 & \text{if base } v \text{ is at site } s \text{ of leaf } C \\ 0 & \text{otherwise} \end{cases}$$

And for a subtree whose root is in $U$:

$$L_{s, \text{low}(RU)}(U = y) = \sum_{q \in \Omega} \left[ P_{xy}(l_B, v_B)L_{s, \text{low}(UB)}(B = q) \right] \times \sum_{v \in \Omega} \left[ P_{yv}(l_C, v_C)L_{s, \text{low}(UC)}(C = v) \right]$$

**Computing the Likelihood When the Model of Evolution Is Reversible**

Reversibility.—When computing the likelihood of a phylogenetic tree, a root $R$ must be specified. If the model is homogeneous and reversible, the process of evolution is stationary: wherever the root is, its base proportions are the same, i.e., they are the equilibrium frequencies of the process, noted $\pi$: $P(R = x) = \pi_x$, (1) can be rewritten:

$$L_s = \sum_{x \in \Omega} \left( \pi_x \sum_{y \in \Omega} \left[ P_{xy}(l_U, v_U) \right] \right)$$

Reversibility means that, averaged over the whole sequence, the flux from one base to another is equal to the flux from this other base back to the first one:

$$\pi_x P_{xy}(l, v) = \pi_y P_{yx}(l, v)$$

**Figure 1.** Example rooted tree for likelihood computation. This tree is composed of a root $R$, an internal node $U$, three other nodes or leaves $A$, $B$, and $C$, and four branches of length $l_A$, $l_B$, $l_C$, $l_U$, and other evolutionary parameters $v_A$, $v_B$, $v_C$, $v_U$. We are here interested in the likelihood of the tree for a single site. The internal node states are unknown and then represented as variables $x$ at node $R$, $y$ at node $U$, $z$ at node $A$, $q$ at node $B$, and $v$ at node $C$. Arrows represent the evolutionary direction, from the root of the tree to its leaves.
Supposing the model used is reversible, as is the case with most current models of DNA sequence evolution, we can rewrite the likelihood in (4) as:

\[ L_s = \sum_{y \in \Omega} \left( \pi_y \sum_{x \in \Omega} \left\{ P_{yx}(l_U, v_U) \times \sum_{z \in \Omega} [P_{xz}(l_A, v_A)L_{s,low}(RA)(A = z)] \right\} \times \sum_{q \in \Omega} [P_{yq}(l_B, v_B)L_{s,low}(UB)(B = q)] \right. \times \sum_{v \in \Omega} [P_{vy}(l_C, v_C)L_{s,low}(UC)(C = v)] \] (5)

Expression (5) can be read as if the root was placed at node \( U \). The root can therefore be placed at any node, on any branch of the tree, a property named “pulley principle” by Felsenstein (1981), and widely used in heuristics to find most likely trees. Considering Figure 1, this makes arrows meaningless.

The possibility to place the root of the tree wherever is needed is thoroughly used in recent heuristics to the problem of the most likely phylogenetic tree. Those heuristics usually explore the space of tree topologies by applying local rearrangements: “nearest neighbor interchange” (NNI) swaps two subtrees around an internal edge (used in PhyML [Guindon and Gascuel, 2003]), “subtree pruning and re-grafting” removes a subtree from the whole tree and places it on another edge (used in RAxML [Stamatakis et al., 2005]), and “tree bisection and reconnection” splits the tree into two subtrees that are rewired by any of their edges. In all these rearrangements, whole subtrees remain fixed: their branches still have the same parameters and their internal topology is unchanged. By defining conditional likelihoods for fixed subtrees, and by placing the root at the rearrangement point, one can avoid much computation when exploring the space of tree topologies. As the root is placed at the rearrangement point, all the conditional likelihoods can be considered as lower from a mathematical point of view since none contains the root.

The most efficient algorithms first compute conditional likelihoods for all subtrees, before they compute an approximate likelihood for topologies obtained with a given sort of rearrangement, using the previously obtained conditional likelihoods. They apply the most promising rearrangements, either all at once (PhyML) or as soon as it is tried (RAxML), optimize evolutionary parameters of the new tree, and eventually start a new round of conditional likelihood calculation and exploration of the space of tree topologies, until convergence.

**Computing the Likelihood of a Tree with a Nonreversible Model**

**Upper conditional likelihoods.**—In the nonreversible case, upper conditional likelihoods can be defined to account for the true root of the tree and the evolutionary directions of the branches.

We define the upper conditional likelihood at branch \( RU \) in the nonreversible case as:

\[ L_{s,upp(RU)}(R = x) = P(R = x) \times \sum_{z \in \Omega} [P_{xz}(l_A, v_A)L_{s,low}(RA)(A = z)] \] (6)

The underlying branches’ upper likelihoods can also be defined recursively:

\[ L_{s,upp(UB)}(U = y) = \sum_{x \in \Omega} [P_{xy}(l_U, v_U)L_{s,upp(RU)}(R = x)] \times \sum_{q \in \Omega} [P_{yq}(l_B, v_B)L_{s,low}(UB)(B = q)] \] (7)

The main difference lies in the incorporation of the root nucleotide frequencies in the definition of the upper conditional likelihoods. This way, the root is not moved around the topology, and the evolutionary direction is conserved.

We now prove that the expression of the tree likelihood is not changed when computed from other nodes of the tree using upper and lower likelihoods.

**Recurrence.**—We show that for any branch, say \( UB \),

\[ L_s = \sum_{y \in \Omega} \left\{ L_{s,upp(UB)}(U = y) \times \sum_{q \in \Omega} [P_{yq}(l_B, v_B)L_{s,low}(UB)(B = q)] \right\} \] (8)

We initialize the recurrence with branch \( RU \):

\[ L_{s,RU} = \sum_{x \in \Omega} \left\{ L_{s,upp(RU)}(R = x) \times \sum_{y \in \Omega} [P_{xy}(l_U, v_U)L_{s,low(UR)}(U = y)] \right\} \]

We expand it:

\[ L_{s,RU} = \sum_{x \in \Omega} \left( P(R = x) \times \sum_{z \in \Omega} [P_{xz}(l_A, v_A)L_{s,low(RA)}(A = z)] \times \sum_{q \in \Omega} [P_{yq}(l_B, v_B)L_{s,low(UB)}(B = q)] \right) \times \sum_{v \in \Omega} [P_{vy}(l_C, v_C)L_{s,low(UC)}(C = v)] \]
With (1):
\[ L_{s,RU} = L_s \]

The likelihood computed on the branch \( RU \) is the same as the one computed at the root. This is also true for the edge \( RA \).

We now suppose we know the likelihood at a branch \( RU \) and are interested in the likelihood at an underlying branch \( UB \).

\[
L_{s,UB} = \sum_{y \in \Omega} \left\{ L_{s,upp(RU)}(U = y) \times \sum_{q \in \Omega} \left[ P_{yy}(l_R, v_R) L_{s,low(RU)}(B = q) \right] \right\} \tag{9}
\]

We expand it, using (7):

\[
L_{s,UB} = \sum_{y \in \Omega} \left\{ \sum_{x \in \Omega} \left[ L_{s,upp(RU)}(R = x) P_{xy}(l_U, v_U) \right] \times \sum_{v \in \Omega} \left[ P_{yu}(l_C, v_C) L_{s,low(UU)}(C = v) \right] \times \sum_{q \in \Omega} \left[ P_{qy}(l_B, v_B) L_{s,low(UB)}(B = q) \right] \right\}
\]

And then rearrange it:

\[
L_{s,UB} = \sum_{x \in \Omega} \left\{ L_{s,upp(RU)}(R = x) \times \sum_{y \in \Omega} \left[ P_{xy}(l_U, v_U) L_{s,low(RU)}(U = y) \right] \right\}
\]

So that we have proved, with (8):

\[
L_{s,UB} = L_{s,RU} = L_s
\]

By recurrence, we have shown that the likelihood value can be computed from any branch of the tree, which, as will be seen, is particularly useful when exploring the space of tree topologies.

**Exploring the Space of Tree Topologies with a Nonreversible Model**

Efficient heuristics explore the space of tree topologies by local rearrangements such as nearest neighbor interchanges (NNIs). In the nonreversible case, evolution proceeds from the root of the tree to its leaves, so one must keep this evolutionary direction unchanged. For this purpose, a distinction is made between the branch on which the root is placed and the others. Figure 2a shows that being able to compute the likelihood from branch \( UB \) permits to define four subtrees whose conditional likelihoods can be used as constants to estimate the likelihoods of the three alternate topologies. In the nonreversible case, one uses upper conditional likelihood for the root-containing subtree and lower likelihoods for all other subtrees. With no loss of generality, NNIs around branch \( UB \) only require exchanges of subtrees having lower conditional likelihoods.

The likelihood of topology 1 (Figure 2a) can be computed with:

\[
L_{s,1} = \sum_{x \in \Omega} \left\{ L_{s,upp(RU)}(R = x) \times \sum_{y \in \Omega} \left[ P_{xy}(l_U, v_U) \times \sum_{v \in \Omega} \left[ P_{yu}(l_C, v_C) L_{s,low(UU)}(C = v) \right] \times \sum_{q \in \Omega} \left[ P_{qy}(l_B, v_B) L_{s,low(UB)}(B = q) \right] \right] \right\}
\]

The likelihoods of topologies 2 and 3 are computed similarly.

In case the internal branch around which NNIs are to be done possesses the root, the situation is slightly different (Fig. 2b). As in the above case, three configurations can be reached through interchanges between subtrees, but here all conditional likelihoods are lower.

The likelihood of the topology 1 (Figure 2b) can be computed as follows:

\[
L_{s,1} = \sum_{x \in \Omega} \left\{ P(R = x) \times \sum_{z \in \Omega} \left[ P_{xz}(l_A, v_A) \times \sum_{i \in \Omega} \left[ P_{zi}(l_F, v_F) L_{s,low(AF)}(F = i) \right] \times \sum_{j \in \Omega} \left[ P_{sj}(l_C, v_C) L_{s,low(AC)}(G = j) \right] \times \sum_{q \in \Omega} \left[ P_{qu}(l_B, v_B) L_{s,low(UB)}(B = q) \right] \right\} \right\}
\]

The likelihoods of topologies 2 and 3 are computed similarly.

It can be shown that by only doing NNIs, the root can be moved throughout the whole tree, to the exception of leaves: as NNIs keep internal branches internal (and external branches external), the root cannot be moved to a leaf.

Thus, the exploration through NNIs of the space of rooted tree topologies with a nonreversible model is as
exhaustive as the exploration of the space of unrooted trees with a reversible model, except that the root cannot go to an external branch.

Once again, and in the same way as computing the likelihood of a tree with a nonreversible model of evolution can be considered as a generalization of the reversible case, the way the space of tree topologies can be explored by NNIs with a nonreversible model of evolution can be seen as a generalization of the reversible case. Overall, nonreversible models of evolution can easily fit into recent heuristics such as PhyML to search for most likely rooted phylogenies.

In the next part, we report a new program built on the algorithmic architecture of PhyML, which explores the space of tree topologies under the nonhomogeneous, nonstationary model of Galtier and Gouy (1998).

**nHPHYML, Adaptation of PhyML Algorithmic Structure to Galtier and Gouy's Model**

We adapted the fast heuristics of PhyML (Guindon and Gascuel, 2003) to Galtier and Gouy’s nonhomogeneous and nonstationary model, and we report here results concerning the ability of the resulting nHPHYML program to explore the space of tree topologies and to estimate the ancestral G+C content.

Galtier and Gouy’s model is particularly variable rich: in addition to common parameters such as branch lengths and transition/transversion ratio, it incorporates different equilibrium G+C contents for each branch and an additional G+C content at the root. This makes it a model containing \(4n - 2\) variables, with \(n\) the number of taxa in the tree: \(2n - 3\) branch lengths, \(2n - 2\) equilibrium G+C contents, the G+C content at the root, the transition/transversion ratio, and an additional parameter defining the position of the root on its branch; i.e., the fraction of the branch length lying on the left side of the root. All these parameters are estimated in the maximum likelihood framework, which leads to a computationally intensive model. For this reason, in all the studies that used this model to find phylogenetic trees (Galtier et al., 1999; Tarrio et al., 2001; Herbeck et al., 2005), no exploration of the space of tree topologies was conducted; the model was simply used to compare a limited set of input phylogenies.
The use of very efficient heuristics to find most likely trees is then mandatory for this kind of model to be used not just as an evaluation tool. PhyML (Guindon and Gascuel, 2003) is such an algorithm that explores the space of tree topologies around an input phylogeny. The adaptation of Galtier and Gouy’s model to the algorithmic structure of PhyML permits for the first time to use this nonstationary, nonhomogeneous model to explore the space of phylogenies with dozens of sequences.

PhyML’s code was deeply modified to produce nhPhyML. nhPhyML starts from a user input–rooted topology: the choice of the root depends upon the user and is unchanged throughout the whole search for the most likely tree, except for its position on its branch. Even if we have shown that NNIs around the root branch could be easily implemented, this has not been done in nhPhyML. Data structures had to be slightly remodelled, as in Galtier and Gouy’s model each branch has its own substitution matrix, and algorithms had to incorporate the fact that the root was fixed, both in the computation of likelihood values for alternate topologies obtained by NNIs (2) and in the computation of conditional likelihoods themselves. Those conditional likelihoods are computed almost as in PhyML, by first a postorder (the original Felsenstein’s pruning algorithm; Felsenstein, 1981) and then a preorder tree traversal, but starting from the root of the tree, whereas in the reversible case the starting point could be any leaf.

Equations (2) and (3) show that lower conditional likelihoods depend, if at a leaf, upon the base observed in the sequence, or, if at an internal node, upon the values of the underlying nodes lower conditional likelihoods. Lower conditional likelihoods can then be obtained by a postorder traversal starting from the root node: the tree is traversed to its leaves, and then the conditional likelihoods of the upper nodes are computed, from the leaves up to the root. On the contrary, upper conditional likelihoods depend both on underlying nodes’ lower conditional likelihoods and on above-lying nodes’ upper conditional likelihoods (Equations (6) and (7)): upper conditional likelihoods can then be computed once lower conditional likelihoods have been computed, and with a preorder tree traversal.

All the parameters of the model are optimized with the Newton-Raphson method (Felsenstein and Churchill, 1996; Galtier and Gouy, 1998). Derivatives are computed analytically except for the shape parameter of the gamma distribution accounting for differences in substitution rates across sites (Yang, 1993) whose derivatives are computed numerically.

The topology is reorganized as in PhyML except that when estimating the approximate likelihood of a given NNI, not only the length but also the equilibrium G+C content of the internal branch around which NNIs are done are optimized.

Results obtained with nhPhyML suggested that the program was prone to getting trapped in local maxima. This prompted us to develop an approximate version of the Galtier and Gouy model, named nhPhyML-Discrete.

We thus adopted a strategy inspired by Foster (2004) and only allowed a limited set of equilibrium frequencies, themselves permanently set to \( \frac{1}{c+1}, \ldots, \frac{c}{c+1} \). Each branch can still have its own equilibrium frequency, by choosing from the few ones available.

Three changes were introduced in the algorithm. First, the user sets the number \( c \) of equilibrium G+C frequencies. Second, before the exploration of the space of tree topologies, each equilibrium frequency is tested for each branch independently from the others, and the branch length is optimized for each equilibrium frequency. The best pairs (equilibrium frequency–branch length) are recorded and ordered according to the gain in likelihood they permit. When all branches have been tried, all the best values are simultaneously used. If the likelihood does not increase, only the first half of them, according to the order previously defined, are applied, until increase. This technique is very similar to the one used in PhyML to optimize branch lengths. Third, the space of tree topologies is explored as in PhyML, except that for each NNI that is tried, all the equilibrium frequencies are tried on the internal branch, and for each one the branch length is optimized.

### Ability of nhPhyML to Explore the Space of Tree Topologies

In order to estimate the ability of nhPhyML to explore the tree topological space, we simulated the evolution of 1000-bp-long sequences according to Galtier and Gouy’s model with a gamma-distributed rate across sites and the rooted version of the trees containing 40 leaves that were used to test PhyML (Guindon and Gascuel, 2003). The ancestral sequence G+C content was uniformly drawn from the interval \([0.2; 0.8]\), and the equilibrium G+C frequencies were uniformly drawn at each node from the interval \([0.1; 0.8]\), but the transition-transversion ratio was kept constant on the whole tree. We then applied various algorithms (neighbor joining, maximum parsimony and maximum likelihood with PhyML) to estimate their ability to find the topologies that had been used to simulate the evolution of the sequences (the “true topologies”) and compared them to nhPhyML.

Among these algorithms, we distinguished programs that do not need a starting topology (like distance-based approaches) from the ones that reorganize a user input tree to explore the space of tree topologies (like nhPhyML). PhyML and the parsimony algorithm can be said to belong to the two classes, as they reorganize a starting topology that can be input a priori by the user or generated by the program itself. Two experiments were then conducted, one in which algorithms that do not need a user input topology were tested upon the simulated sequences (Fig. 3, white bars), and one in which algorithms that can run starting from a user input topology were compared (Fig. 3, grey bars). For this second experiment, input topologies were obtained by perturbing the “true topologies” by a number of NNIs uniformly drawn from \([5; 20]\), while making sure that the ingroup and the outgroup were not melted. This additional condition is necessary as nhPhyML is not able to question
Figure 3. Efficiency of various methods in reconstructing a phylogeny in nonhomogeneous conditions. White bars: Results obtained by phylogenetic methods that do not start from a user input topology. Grey bars: Results obtained by phylogenetic methods that start from perturbed input topologies. These results were obtained with the same 2000 different topologies taken from the PhyML test set. Error bars represent standard deviations. Asterisks are displayed where Student paired t-tests are significant at the 1% level.

The position of the root when exploring the space of tree topologies.

To estimate the efficiency of a method, Robinson and Foulds' (R&F; Robinson and Foulds, 1979) average distances between the true topologies and the reconstructed ones were computed with the PHYLIP package (Felsenstein, 1989). Results are given in Figure 3.

For both PhyML (version 2.4.4, under the TN93 model [Tamura and Nei, 1993]) and nhPhyML, reconstruction was made using a gamma law with eight categories to account for across-site rate variation; parameter $\alpha$, transition/transversion ratio, and the other parameters were estimated by the programs. PhyML and the parsimony method as implemented in PAUP* (Swofford, 2003) were both used from their built-in starting topology and from the perturbed input topologies. The neighbor-joining algorithm was applied to pairwise distances estimated under HKY85 (Hasegawa et al., 1985), LogDet (Lake, 1994; Lockhart et al., 1994), GG95 (Galtier and Gouy, 1995), and transversions-only observed divergence distances.

Figure 3 shows that PhyML is more efficient at finding good topologies than parsimony, which also has better results than distance methods. It is surprising to note that the transversions-only observed divergence and the GG95 distances perform worse than the HKY85 distance, because these methods were devised to be resistant to G+C content biases. However, as expected, the LogDet distance provides better results than the HKY85 distance.

Figure 3 (grey bars) compares the efficiencies of parsimony, PhyML, and nhPhyML methods. The average distance between the rearranged input topologies and the true topologies is shown. All the methods tested are able to explore the space of tree topologies to find better trees than the input ones. Results obtained by PhyML from the rearranged input phylogenies are better than when PhyML departs from its own starting topology (a distance-based tree). As the rearranged topologies are further away from the true topologies than the distance-based trees, this comes from the fact that distance trees fail on subtrees also difficult to solve for the maximum likelihood method, whereas the rearranged topologies can be perturbed at the level of subtrees whose solution is trivial.

It appears that PhyML is able to find better trees than parsimony and surprisingly also better trees than...
nPhyML (Student unilateral paired t-test, P-value <1 x 10^-10). This might be due to the large number of parameters: nPhyML has 2 x n - 2 additional parameters when compared to PhyML, because an equilibrium G+C content is associated to each branch. This might result in a likelihood surface with lots of local maxima, in which the algorithm would get trapped.

The comparison of the likelihood values found by nPhyML when launched from the rearranged topologies to the likelihood values computed on the true topologies comforts us in this hypothesis: the log-likelihood of the true trees is on average 62.9 points higher than the log-likelihood of the trees found by nPhyML, which have a better likelihood than the true trees in only 830 cases out of 2000. As a comparison, PhyML finds topologies with a 1.2-point higher log-likelihood score than the true topologies and finds topologies with better likelihoods than the true ones in 1562 cases out of 2000. This means that nPhyML fails to correctly explore the space of tree topologies, because it gets trapped in local maxima and does not get to the real maximum. Being particularly parameter rich, it seems that nPhyML can fit nearly any topology by taking advantage of its numerous parameters.

An approximate and less flexible model was developed and was named nPhyML-Discrete. This nonhomogeneous model still has the same number of free parameters as nPhyML, as each branch can have a particular equilibrium frequency, but is also much less "flexible," because these equilibrium frequencies are constrained to be in the limited set defined by the user. These constraints have a positive impact on the results of the algorithm. These are shown in Figure 3 for sets of 1 to 4 equilibrium frequencies.

nPhyML-Discrete shows a better topological accuracy than PhyML even when using only one equilibrium frequency: this can be due either to the fact that the root base distribution is estimated in nPhyML-Discrete and not in PhyML, or to the fact that the ingroup and the outgroup cannot be swapped in nPhyML-Discrete, thereby avoiding the exploration of unreasonable tree topologies contrary to PhyML. Increasing the number of equilibrium frequencies further increases nPhyML-Discrete's topological accuracy, but this tendency quickly reverses, as using 3 equilibrium frequencies yields better results than 4 equilibrium frequencies (though the unilateral paired Student t-test is not significant). When further increasing the number of equilibrium frequencies, the topological accuracy continues dropping, with, for instance, an average distance to the true topologies of 2.23 for 10 equilibrium frequencies (data not shown), not better than when using only 1 equilibrium frequency (2.18, Student unilateral paired t-test: P-value: 0.054). Overall, it seems that using 3 equilibrium frequencies might be a good choice, as it gets the best topological accuracy on the simulations, which is the same performance as with 2 equilibrium frequencies, but with a lower standard deviation.

When used with 3 G+C equilibrium frequencies, the algorithm finds topologies closer to the true ones than PhyML (nPhyML-Discrete: 2.09, PhyML: 2.49, Student unilateral paired t-test, P-value <1 x 10^-10). This also has an impact on the risk of getting trapped in local optima: nPhyML-Discrete finds topologies that have a log-likelihood on average 37.5 points higher than the true topology log-likelihoods, which is better than nPhyML. Moreover, it appears that nPhyML-Discrete finds trees with better likelihoods than the true tree in 1768 cases out of 2000. Overall, it seems that nPhyML-Discrete shows a performance as good as PhyML's one, with a stronger variation in log-likelihood, which might hint for a stronger discriminating power. Finally, this approximation has a great impact on the computational speed, nPhyML used on average 40 min 42 s to give its results while nPhyML-Discrete only needed 20 min 43 s: it is faster to choose among a limited set of equilibrium frequencies than to optimize the value of a continuous parameter.

Figure 4 shows that nPhyML-Discrete is, as PhyML and parsimony, nearly insensitive to the distance of its input phylogeny to the true one. On the contrary, nPhyML does not seem to be able to cope with distant topologies, which is in agreement with the results above.

**Estimation of Root G+C Content**

The ancestral G+C content is a parameter of the model in itself. We conducted tests to check the ability of the program to estimate this parameter on trees containing 40 leaves, either from the true phylogenies or from the phylogenies found by nPhyML and nPhyML-Discrete when it was input perturbed topologies. nPhyML-Discrete results are shown Figure 5 for 3 equilibrium frequencies.

Whether it is estimated from the true phylogenies, or from the phylogenies found by nPhyML or nPhyML-Discrete (Fig. 5), the ancestral G+C content is well estimated. The correlation coefficient between estimated and expected G+C contents is above 0.99 in all cases. Interestingly, results do not depend upon the number of equilibrium frequencies: performances are highly similar whether we use only 1 equilibrium frequency or whether we use nPhyML. The average of the squared differences between the estimated and the true values is ∼0.000282 when inferred from the true phylogeny, ∼0.000292 when inferred from the phylogenies found by nPhyML, and ∼0.0003423 when inferred from the phylogenies found by nPhyML-Discrete with 3 equilibrium frequencies from the phylogenies it has found. It is interesting to note that even if nPhyML-Discrete does not model evolution as precisely as nPhyML, being limited in its choice of equilibrium frequencies, it can still provide very good estimates of the ancestral G+C contents, even from topologies that are not the true ones.

Overall it appears that the limitation of the number and values of equilibrium frequencies has been very successful, permitting to increase the ability of the algorithm to explore the space of tree topologies while retaining the capacity to estimate ancestral G+C content.
FIGURE 4. Ability of phylogenetic methods to explore the space of tree topologies in nonhomogeneous conditions. As the distance from the input phylogenies to the true topologies increases, results obtained by nhPhyML get worse, but nhPhyML-Discrete seems to be less dependent upon the input topologies.

FIGURE 5. Ability of nhPhyML-Discrete to estimate ancestral G+C contents. nhPhyML-Discrete was used to estimate ancestral G+C contents on sequences simulated on 2000 different topologies from the PhyML test set (see text) from the phylogenies found by the program itself, with 3 equilibrium frequency categories. In each experiment, parameter $\alpha$ that accounts for across-site rate variation using an 8-category discretized gamma distribution, transition/transversion ratio, and the other parameters were estimated by maximizing the likelihood.

ESTIMATION OF THE UNIVERSAL PHYLOGENY AND OF THE ANCESTRAL G+C CONTENT

The ability of nhPhyML-Discrete to explore the space of tree topologies and to estimate the ancestral G+C content has then been demonstrated and can be applied to real data.

As rRNA stem G+C contents are known to be correlated to the optimal growth temperatures in Bacteria and Archaea (Galtier and Lobry, 1997), these genes are especially good candidates for an analysis using Galtier and Gouy's model (Galtier and Gouy, 1998), and hence nhPhyML-Discrete. Therefore, we reiterated the analysis of Galtier et al. (1999), improving the taxonomic sampling and benefiting from the heuristics of PhyML to better scan the space of tree topologies.

The analysis was divided in two steps: first we searched with nhPhyML-Discrete for the most likely topology for which complete rRNA genes and the model plead, and then, using the best topology we could find and the stem portion of rRNA sequences, we estimated the G+C content of the last universal common ancestor (LUCA) with nhPhyML.

Phylogeny Estimation

Small and large rRNA subunit sequences were downloaded from the European Ribosomal RNA database...
branch lengths. None of these 18 topologies were identified as being significantly more likely than the others (AU test P-value > 0.05) among the starting phylogenies, among which 9 had identical topologies.

The resulting alignment contained 2924 sites, with G+C contents ranging from 43% to 71%, which highlights the need for models robust to compositional biases.

Even if the ability of nhPhyML-Discrete to explore the space of tree topologies appears as good as PhyML's, it is wise to run the program from various starting topologies to diminish the risk of getting trapped in local optima. Moreover, as the process of evolution is not reversible, the position of the root influences the likelihood value. For this reason, it was decided to build various topologies with distance-based and parsimony methods (as implemented in Phylo.Win, Galtier et al., 1996), and then to try three different rootings for each of these topologies. The trees were rooted either on the branch leading to the Archaea, on the branch leading to the Bacteria, or on the branch leading to the Eukaryota. This produced 57 starting phylogenies, among which 9 had identical topologies.

Results obtained by nhPhyML-Discrete on these trees were then analyzed using CONSEL (Shimodaira and Hasegawa, 2001) and Treedist (PHYLIP package; Felsenstein, 1989). Eighteen phylogenies were found to be significantly more likely than the others (AU test P-value < 5%) among the 55 different topologies found, which shows that even if nhPhyML-Discrete showed good capabilities to explore the space of tree topologies, on real cases, with many sequences, the algorithm can get trapped in a local maximum, even when starting from two phylogenies with the same topologies but different branch lengths. None of these 18 topologies were identical, so the majority rule (extended) consensus tree obtained from these trees was built (Fig. 6) using Consense from the PHYLIP package (Felsenstein, 1989).

The Tree of Life

Though we do not believe that a two-gene phylogeny can clarify the Tree of Life, we think the analysis of rRNAs using a nonhomogeneous model might bear some interesting insights.

Most great clades are found monophyletic; e.g., Proteobacteria, Metazoa, Crenarchaeota. In the Bacteria kingdom, Firmicutes and Clostridia are found associated in all trees. Planctomycetes are found monophyletic and are associated with Chlamydiales, but the basal position is not clear (as in Brochier and Philippe, 2002). Slowly evolving sites were identified by getting rid of fastest evolving positions on the small subunit rRNA gene and is not found when coping with compositional heterogeneities on LSU and SSU rRNA genes. Instead, hyperthermophilic Bacteria are found at the root of the bacterial clade, as in first rRNA phylogenies (Woese, 1987).

Ancestral G+C Content Estimate

The consensus universal phylogeny found using nhPhyML-Discrete was used to infer the ancestral G+C contents of the small- and large-subunit rRNA genes. The inference was performed using only slowly evolving stem regions for two reasons. First, only rRNA stems, that is, the fraction of the rRNA molecule that folds as double helices, have a G+C content strongly correlated with optimal growth temperature (Galtier and Lobry, 1997). Second, Gowri-Shankar and Rattray (2006) have recently shown that the ancestral G+C content estimate obtained with the Galtier and Gouy model was biased towards the G+C content at slowly evolving sites and that equilibrium frequencies were biased towards fast-evolving sites. Correlations between interacting sites of the helices were not taken into consideration.

Stem regions were identified using the following procedure. The rRNA alignment downloaded from the European Ribosomal RNA database (Wuyts et al., 2004) was extended to the present sample of 92 sequences by manually aligning missing sequences using SeaView (Galtier et al., 1996). A total of 1896 sites were predicted in stems in Escherichia coli, Archaeoglobus fulgidus, and Schizosaccharomyces pombe. Slowly evolving sites were identified by using the COE program (Dutheil et al., 2005) from the Bio++ package (Dutheil et al., 2006) and selecting sites predicted to undergo on average less than 0.1 substitution per branch under a HKY85 model with an 8-class discretized gamma law to model rate heterogeneity. Six hundred seventy-eight slowly evolving stem sites were finally retained.

The correlation between rRNA stem slowly evolving sites G+C content and optimal growth temperature (Topt) is high for both Bacteria (0.815) and Archaea (0.953), which is in agreement with Galtier and Lobry (1997). Because ancestral G+C content inferences are known to be robust with respect to the tree topology (Galtier et al., 1999; and Fig. 5 herein), those estimates are expected not to depend strongly on the input phylogeny.

Because the location of the root of the universal tree is not currently known (Brown and Doolittle, 1997; Forterre and Philippe, 1999), the likelihoods of all three possible rootings, that is, on the branch leading to each one of the three domains, were computed using nhPhyML on the slowly evolving stem sites. We chose not to use the discrete version of nhPhyML in order to avoid any bias that might arise from the fact that equilibrium frequencies are set to a priori values. For instance, with three G+C categories, the equilibrium frequencies are set to 0.25, 0.50, or 0.75, which may not model the evolution of the slowly evolving stem sites appropriately, given that their G+C content range from 0.52 for Entamoeba histolytica to 0.83 for Methanopyrus kandleri. Four rate categories were used to model rate heterogeneity, and the parameter α of the gamma distribution and the transition/transversion ratio were optimized by maximizing the likelihood. Since studies (Huelsenbeck et al., 2002; Yap and Speed, 2005) have shown that there may be information in extant sequences for the identification of the root position of a phylogenetic tree using non-reversible models of evolution, we compared all three likelihoods to see which root the model and the data were predicting. We used the slowly evolving stem sites, as they are expected to be less prone to saturation problems. The
FIGURE 6. Consensus tree obtained from the 18 significantly most likely trees obtained by nhPhyML-Discrete. The tree was built as described in the text. Dashed lines represent branches that were not found in all the 18 most likely trees. The alignment contained 2924 sites. Group names are in agreement with the NCBI taxonomy. Sequence G+C contents range from 43% to 71%, which highlights the need to use models robust to compositional biases.
highest likelihood was obtained by the bacterial branch rooting (log-likelihood: -15,139.94, ancestral G+C content: 71.8%), whereas the least realistic rooting was found on the archaeal branch (log-likelihood: -15,147.22, ancestral G+C content: 75.1%) and could be rejected using CONSEL (Shimodaira and Hasegawa, 2001) (AU test P-value <5%). Hence, we chose to discard the archaeal rooting from subsequent analyses. There was no significant difference between the bacterial and the eukaryotic rootings (log-likelihood: -15,141.61, ancestral G+C content: 69.2%). It seems interesting to note that the longest branch (the eukaryotic branch) was not found to provide the most likely rooting; the model is then able to find a signal that is independent from the evolutionary distance.

To estimate the accuracy of the estimation of ancestral G+C contents, we computed the likelihoods obtained when setting the ancestral G+C content to various values, between 55% and 85%, and compared the results using CONSEL. Ancestral G+C contents with AU test P-values higher than 5% were considered to be in the confidence interval. We found that, when rooting in the eukaryotic branch, ancestral G+C contents ranging from 63% to 75% could not be rejected, whereas when rooting in the bacterial branch, the confidence interval was [67%; 76%]. The larger confidence interval found for the eukaryotic rooting might be explained by the fact that this branch is considerably longer than the bacterial one: the extra length of the eukaryotic branch may provide more latitude to accommodate nonoptimal ancestral G+C contents.

Inferred ancestral G+C contents (Fig. 7) suggest a mesophilic (optimal growth temperature below 60°C) to thermophilic (optimal growth temperature between 60°C and 80°C) LUCA, in agreement with Galtier et al. (1999). Interestingly, both confidence intervals do not contain any value that seem to favour a hyperthermophilic LUCA. As a consequence, it appears that reducing site rate heterogeneity to avoid the bias put forth by Gowri-Shankar and Rattray (2006) does not contradict Galtier et al.’s conclusion.

CONCLUSION

In this article, we have shown that by reorganizing the way likelihood is computed, one can efficiently explore the space of tree topologies with a nonreversible model of evolution. We modified the PhyML algorithm
(Guindon and Gascuel, 2003) to cope with the nonhomogeneous, nonstationary model of Galtier and Gouy (1998) and tested its abilities to find the right topology and to estimate the G+C content at the root. An approximate model was also tested, which showed good performance in both tree topological space exploration and ancestral G+C content estimation. We eventually used the program to estimate the topology of the Tree of Life from rRNA sequences and to estimate the ancestral stem G+C content by only selecting slowly evolving sites. The results agree with the ones obtained by Galtier and Lobry (1999) and support a nonhyperthermophilic last universal common ancestor.

Genome sequences vary widely in their composition between species. Therefore, when building a phylogenetic tree from such heterogeneous data, it is important to use a method robust to compositional biases. Nonhomogeneous models of evolution are particularly suitable, but their nonreversibility discarded them from most general phylogeny packages and prevented their use in large scale analyses. This work renders nonreversible models of evolution useful for phylogeny reconstruction, which considerably broadens the range of available models and opens new opportunities for models explicitly dealing with compositional biases.

ACKNOWLEDGEMENTS

nhPhyML is available as a LINUX executable file at http://pbil.univ-lyon1.fr/software/nhphyml/ and as source code upon request. We want to thank the reviewers and Olivier Gascuel for their constructive remarks, which resulted in considerable improvement in the quality of the manuscript. This work was supported by Action Concertée Incitative IMPBIO. We thank the Centre de Calcul de l’IN2P3 for providing computer resources. Bastien Boussau acknowledges a PhD scholarship from the Centre National de la Recherche Scientifique. We also thank Mathilde Paris for fruitful discussions.

REFERENCES


Associate Editor: Olivier Gascuel