A Stochastic Local Search Algorithm for Distance-Based Phylogeny Reconstruction

Francesca Tria,*,[1] Emanuele Caglioti,[2] Vittorio Loreto,[1,3] and Andrea Pagnani[1]

[1]Institute for Scientific Interchange, Torino, Italy
[2]Dipartimento di Matematica, Sapienza Università di Roma, Roma, Italy
[3]Dipartimento di Fisica, Sapienza Università di Roma, Roma, Italy

***Corresponding author:** E-mail: tria@isi.it.

Associate editor: Alexei Drummond

Abstract

In many interesting cases, the reconstruction of a correct phylogeny is blurred by high mutation rates and/or horizontal transfer events. As a consequence, a divergence arises between the true evolutionary distances and the differences between pairs of taxa as inferred from available data, making the phylogenetic reconstruction a challenging problem. Mathematically, this divergence translates in a loss of additivity of the actual distances between taxa. In distance-based reconstruction methods, two properties of additive distances have been extensively exploited as antagonist criteria to drive phylogeny reconstruction: On the one hand, a local property of quartets, that is, sets of four taxa in a tree, the four-points condition; on the other hand, a recently proposed formula that allows to write the tree length as a function of the distances between taxa, Pauplin's formula. Here, we introduce a new reconstruction scheme that exploits in a unified framework both the four-points condition and the Pauplin's formula. We propose, in particular, a new general class of distance-based Stochastic Local Search algorithms, which reduces in a limit case to the minimization of Pauplin's length. When tested on artificially generated phylogenies, our Stochastic Big-Quartet Swapping algorithmic scheme significantly outperforms state-of-art distance-based algorithms in cases of deviation from additivity due to high rate of back mutations. A significant improvement is also observed with respect to the state-of-art algorithms in the case of high rate of horizontal transfer.

Key words: phylogeny, stochastic methods, noise and horizontal transfer, trees.

Introduction

Phylogenetic methods have recently been rediscovered in several interesting areas among which immunodynamics, epidemiology, and many branches of evolutionary dynamics. The reconstruction of phylogenetic trees belongs to a general class of inverse problems whose relevance is now well established in many different disciplines ranging from biology to linguistics and social sciences (Gray and Atkinson 2003; Lazer et al. 2009; Liu et al. 2009; Pybus and Rambaut 2009). In a generic inverse problem, one is given a set of data and has to infer the most likely dynamical evolution process that presumably produced the given data set. The relevance of inverse problems has been certainly triggered by the fast progress in data-revealing technologies. In molecular biology, for instance, a great amount of genomes data are available thanks to the new high-throughput methods for genome analysis (Ragoussis 2009). In historical linguistics (Renfrew et al. 2000), a remarkable effort has been recently done for the compilation of corpora of homologous features (lexical, phonological, syntactic) or characters for many different languages.

Although phylogenetic reconstruction is not a novel topic, dealing with not purely tree-like processes and identifying the possible sources of nonadditivity and their effects in a given data set is still an open and challenging problem (Felsenstein 2004; Gascuel 2007).

Here, we focus on distance-based methods (Cavalli-Sforza and Edwards 1967; Fitch and Margoliash 1967) and investigate how deviations from additivity affect their performances. In distance-based methods, only distances between leaves are considered, and all the information possibly encoded in the combinatorial structure of the character states is lost. Despite their simplicity, distance-based methods are still widely used thanks to their computational efficiency, but a solid theoretical understanding on the limitation of their applicability is still lacking. One of the most popular distance-based reconstruction algorithms, Neighbor-Joining (NJ; Saitou and Nei 1987), was proposed in the late 1980s, but it is only recently that its theoretical background was put on a more solid basis (Atteson 1997; Gascuel and Steel 2006; Mihaescu et al. 2007). Another step toward a better understanding of distance-based methods was obtained thanks to an interesting property of additive distances, Pauplin's formula (Pauplin 2000). This property has been used in the formulation of a novel algorithmic strategy with improved performances (FastME; Desper and Gascuel 2002). In parallel, another fundamental property of additive trees, the "four-points" condition (Buneman 1971, 1974; Gusfield 1997), has been extensively exploited in distance-based phylogenetic reconstruction methods (Erdös et al. 1998; Bruno et al. 2000; Snir et al. 2008). Both the Pauplin's formula and the four-points condition will be discussed in details below.

Here, we propose a new approach that combines the four-points condition and Pauplin's formula in a Stochastic Local Search (SLS) scheme that we name Stochastic

Big-Quartet Swapping (SBiX) algorithm. SLS (Hoos and Stützle 2005) algorithms transverse the search space of a given problem in a systematic way, allowing for a sampling of low-cost configurations. SLS algorithms start from a randomly chosen initial configuration. Subsequently, the elementary step connects neighboring configurations. Each move is determined by a decision based on local knowledge only. Typically, the decision is taken combining, with a given a priori probability, a greedy step (i.e., a step that reduces the local cost contribution) with a random one, where the local cost is not taken into account. SLS algorithms have been widely used in solving complex combinatorial optimization problems such as Satisfiability, Coloring, MAX-SAT, and Traveling Salesman Problem (Hoos and Stützle 2005).

At the heart of our new algorithmic scheme (named SBiX), there is the notion of "quartet frustration," a quantitative measure of how good a given configuration is, in the space of trees. Following a concept already introduced in Snir et al. (2008), we weigh the different quartets according to their length in order to reduce the effect of those, which are more likely to undergo double mutations. The strength of our approach comes from a combination of this strategy with a Pauplin's-like one, weighing each quartet according to a purely topological property.

We tested the performances of the proposed reconstruction algorithm, that is, the ability to reconstruct the true topology, in the presence of high levels of deviation from additivity due to both horizontal transfer and back-mutation processes. We use both a very simple model to generate artificial phylogenies of binary sequences, and the more realistic Kimura two-parameters model, considering sequences with q-state sites, where $q = 4$. We have evidence that the performance of our algorithm rely neither on the particular evolutionary process giving rise to the phylogeny nor on the particular representation of the taxa. We find that when the lack of additivity arises from high mutation rates (and consequently high probability of back mutations), our algorithm significantly outperforms the state-of-art distance-based algorithms. When the lack of additivity arises from high rate of horizontal transfer events, our algorithm performs better than the algorithms we considered as competing ones.

We show results both for a greedy and for a simulated annealing–like strategies of our algorithm, the former being significantly faster than the latter and with comparable performances. Defining N as the number of taxa, the SBiX algorithm has a complexity of $O(N^4)$, which is higher that the one of the distance-based algorithms used as competitors for comparing the performances. Nevertheless, the prefactor of the greedy version is so low that our algorithm is fast enough to reconstruct large phylogenies, for example, of a few thousands taxa, in a time remarkably slower with respect to any character-based reconstruction algorithm. A comparison of the running time and the performances of our algorithm with a popular character-based one, MrBayes (Huelsenbeck and Ronquist 2001), is reported, respectively, in the Supplementary Material online and in the Results section. We show that although MrBayes slightly outperforms SBiX, the running times of MrBayes greatly exceed ours, becoming comparable when the reconstruction becomes in practice unfeasible, that is, for running times of the order of some thousand of years, and number of taxa of the order of 100,000. A C implementation of the algorithm is available upon request.

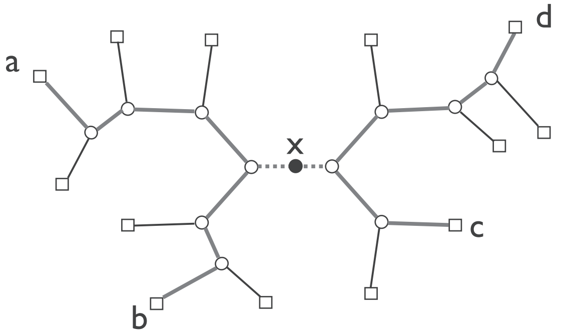

Fig. 1. Quartet definition. The quartet a, b, c, d induces an internal edge x that divides the tree in two parts. All paths joining any pair of sites sitting on opposite parts of the tree pass through x.

Methods

Additivity and the Four-Points Condition

An $N \times N$ distance matrix $\mathscr{D}$ is said to be additive if it can be constructed as the sum of a tree's branches lengths. Two fundamental and widely used properties of additive distances are Pauplin's formula (Pauplin 2000) and the four-points condition (Buneman 1971, 1974; Gusfield 1997), satisfied by each possible group of four within N taxa. For the sake of clarity, we recall them here.

Given any quadruplet of taxa a, b, c, d, let $D_1 = \mathscr{D}(a, b) + \mathscr{D}(c, d)$, $D_2 = \mathscr{D}(a, c) + \mathscr{D}(b, d)$, and $D_3 = \mathscr{D}(a, d) + \mathscr{D}(b, c)$ be the three possible pairs of distances between the four taxa. A matrix $\mathscr{D}$ is additive if and only if $D_1 < D_2 = D_3$ or $D_2 < D_1 = D_3$ or $D_3 < D_1 = D_2$ (It is important to remark here that the four-points condition is an equivalent definition of additivity. That is, a distance matrix is additive if and only if the four-points condition is satisfied.). When considering experimental data, additivity is almost always violated. In order to set up a robust method for phylogeny reconstruction based on the notion of additivity, we need to relax the four-points condition and to quantify violations in a suitable way. We define at this aim a "weak four-points" condition: For any four taxa a, b, c, d such that a, b are on one side of the tree and c, d on the other (as in fig. 1), the quartet $(ab{:}cd)$ is said to satisfy the weak four-points condition if $D_1 = \min(D_1, D_2, D_3)$ (where D_1, D_2, and D_3 are defined as above). It is easy to prove that if the distance matrix $\mathscr{D}$ is additive, a unique tree exists in which all quartets satisfy the weak four-points condition and this tree is the correct one. Many algorithms have been proposed that exploit this weak four-points condition, one of the most promising being, for instance, the short-quartet method (Erdös et al. 1998; Snir et al. 2008).

Pauplin's Distance

Another remarkable property of an additive tree is the possibility to compute its total length L, defined as the sum of all its branches lengths, through a formula, due to Pauplin (Pauplin 2000), that only uses distances between taxa:

$$L_P = \sum_{a<b} 2^{-t(a,b)} \mathscr{D}(a,b), \qquad (1)$$

where $t(a,b)$ is the number of nodes on the path connecting a and b, that is, their "topological distance." For additive trees, $L \equiv L_P$. Even when the four-points condition is no longer satisfied, L_P is a particularly good approximation for the tree length (Desper and Gascuel 2002) and it is recognized that for distance matrices sufficiently "close" to additivity (Atteson 1997), the correct phylogeny minimizes L_P (Mihaescu et al. 2007; Bordewich et al. 2009). This principle is used in an implicit way in NJ (Saitou and Nei 1987) and more explicitly in a new generation distance-based algorithm, FastME (Desper and Gascuel 2002). When departure from additivity is too strong, L_P is no longer a good functional to minimize in order to recover the correct tree.

Violations of Additivity

Violations of additivity can arise both from experimental noise and from properties of the evolutionary process the data come from. We here consider two of the main sources of violations of the latter type, which can either occur together or singularly. 1) "Back mutation": in particularly long phylogenies, here the timescale being set by the mutation rate, a single character may experience multiple mutations; in this case, the distances between taxa are no longer proportional to their evolutionary distances. In the following, we will use the expression back mutation as synonymous of multiple mutation on the same site and 2) "Horizontal transfer": the reconstruction of a phylogenetic tree lies on the the assumption that information flows "vertically" from ancestors to offsprings. However, in many processes, information flows also horizontally. Horizontal (or lateral) gene transfers (Simonson et al. 2005) are well-known confounding factors for a correct phylogenetic inference.

The SBiX Algorithm

Here, we describe the structure of our SBiX algorithm. As already mentioned, this algorithm crucially exploits both the weak four-points condition and the Pauplin's distance. It features a larger robustness with respect to violations of additivity if compared with algorithms based separately on the weak four-points condition or on the minimization of Pauplin's length.

The general structure of the SBiX algorithm is as follows:

1. start with a tree topology for the given set of taxa,
2. update the tree topology by local elementary rearrangements through Nearest Neighbor Interchange (or Big-Quartet Swapping, see fig. 2), and
3. repeat point 2 till convergence is reached.

We analyze the three points in details in the following, where we consider both a simulated annealing–like and a greedy strategies:

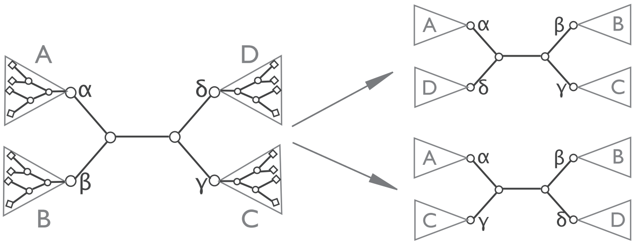

FIG. 2. Nearest Neighbor Interchange. Any internal edge defines the four subtrees A, B, C, D rooted, respectively, on $\alpha, \beta, \gamma, \delta$. Here, the initial reference configuration $((A,B),(C,D))$ displayed on the left leads to two possible rewiring: 1) swap the pair $B \leftrightarrow D$ (right-upper panel) and 2) swap the pair $B \leftrightarrow C$ (right-lower panel).

1. In the simulated annealing–like version of the algorithm, we start with a random topology. In the greedy version, we search for a local minimum (which can eventually be the global one), so it is important to start with a meaningful topology. We tested the algorithm starting both with the FastME and with the NJ reconstructed topologies.
2. We sequentially consider all the internal edges of the present topology. Each internal edge defines four subtrees (see fig. 2), say A, B, C, D, rooted, respectively, on the four internal nodes $\alpha, \beta, \gamma, \delta$. Referring to figure 2, let $((A,B),(C,D))$ being the initial configuration. We randomly choose one of the two possible local rewirings: 1) swap the pair $B \leftrightarrow D$ getting the configuration $((A,D),(B,C))$ and 2) swap the pair $B \leftrightarrow C$ getting the configuration $((A,C),(B,D))$. In the simulated-annealing version, the new configuration is accepted with probability 1 if $\Delta E < 0$ and a probability proportional to the statistical weight $e^{-\beta \mathrm{sign}(\Delta E)}$ if $\Delta E > 0$, where β is an inverse temperature-like parameter that is set by a simulated annealing–like (Kirkpatrick et al. 1983) strategy (see point 3) and ΔE is the difference of the two configurations local costs: For case (1), $\Delta E = E_{((A,D),(B,C))} - E_{((A,B),(C,D))}$, whereas for case (2), $\Delta E = E_{((A,C),(B,D))} - E_{((A,B),(C,D))}$. In the greedy (or zero temperature) version, the new configuration is accepted if and only if $\Delta E < 0$.
3. In the simulated-annealing version, we iterate point 2 starting with $\beta = 0$ and increasing it at a constant rate at each sweep (where we call "sweep" an update of all the internal edges selected in a random permutation, so that each internal edge is selected once and only once, with a different order, for each sweep) until convergence is reached, that is, until the algorithm gets stuck in a fixed topology. In the greedy version, we iterate point 2 until the algorithm gets stuck in a fixed topology.

Let us stress that we use the term "simulated annealing" in a loose sense, because, as we shall discuss in the Remarks on the Cost Definition section, we are not going to minimize a global cost functional.

In order to define the configurational local cost, say $E_{((A,B),(C,D))}$, we consider all the quartets $(ab{:}cd)$ such that $a \in A$ (a is a taxa of the subtree A), $b \in B$, $c \in C$, and

$d \in D$. For each quartet, we define the “quartet frustration” as follows:

$$f_{(ab:cd)} = \max\left(0, \frac{D_1 - \min(D_2, D_3)}{(D_1 + \min(D_2, D_3))^k}\right), \quad (2)$$

where D_1, D_2, and D_3 are the sums of distances already defined ($D_1 = \mathscr{D}(a,b)+\mathscr{D}(c,d), D_2 = \mathscr{D}(a,c)+\mathscr{D}(b,d)$, and $D_3 = \mathscr{D}(a,d) + \mathscr{D}(b,c)$). The normalization factor in the right-hand side of equation (2), as already pointed out in the Introduction section, gives a smaller weight to longer distances, typically affected by noise and recombination. The parameter fixing strategy for the exponent k will be discussed in the Remarks on the Cost Definition section.

The cost $E_{((A,B),(C,D))}$ of the configuration $((A,B),(C,D))$ is thus defined as the sum of the costs (“frustrations”) of all the considered quartets, each weighted with a factor borrowed from the Pauplin’s formula:

$$E_{((A,B),(C,D))} = \sum_{(ab:cd)} f_{(ab:cd)} 2^{-t(a,\alpha)-t(b,\beta)-t(c,\gamma)-t(d,\delta)}, \quad (3)$$

where $t(a,\alpha)$ is the topological distance between the taxa a and the internal node α and analogously for the other taxa. We define the topological distance between a leaf and an internal node as the number of nodes in the path connecting them, “including the considered internal node”: This is set in order to satisfy $\sum_a 2^{-t(a,\alpha)} = 1$ in each subtree.

Remarks on the Cost Definition

When $k = 0$ in equation (2), our procedure is equivalent to the minimization of Pauplin’s length (the proof of this statement will be discussed in the Appendix A). On the other hand, if Pauplin’s weights $2^{-t(a,\alpha)}$, $2^{-t(b,\beta)}$, $2^{-t(c,\gamma)}$, and $2^{-t(d,\delta)}$ were absent, the difference in local costs between two configurations would be equal to the variation of a global cost, defined as $E = \sum_{(abcd)} f_{(ab:cd)}$. Here, the sum defining the cost, $\sum_{(abcd)}$, is running over all the quartets of the tree and not only on the quartets compatible with the subtrees A, B, C, D. We will refer to our algorithm with this form of the cost configuration as the “normalized quartets” (NQ) method.

Conversely, when one takes the complete form of the local cost as defined in equation (3), with $k > 0$ in equation (2), the local cost differences do not correspond to any global cost difference (the proof of this statement will be discussed in Appendix B). It is, however, an open question whether a global functional can be defined whose variation between each pair of configurations is compatible with the sign of our local cost difference.

The complexity of a sweep of our algorithm (i.e., N configurations updates, where N is the number of leaves in the tree has a leading term $O(N^4)$ (The number of quartets to be considered when updating all the edges of the tree in case of a perfectly balanced tree reads

$$\mathscr{N} = \frac{85}{5376}N^4 - \frac{N^2}{3} + \frac{4}{7}N. \quad (4)$$

Note that the above formula for $\mathscr{N}$ is an upper bound for more general topologies.). We show the results of numerical simulations for the running time of the greedy version of our algorithm in the Supplementary Material online. Despite the $O(N^4)$ complexity of the SBiX algorithm, the greedy version has an extremely low prefactor, making the algorithm suitable for trees with a large number of taxa (see Supplementary Material online for details).

Results

Artificial Phylogenies

To test the performances of our algorithm, we consider artificially generated phylogenies, following one of the simplest evolutionary model that takes into account both mutational events and horizontal transfer. Each taxon is represented by a binary sequence of length l. We start with one sequence, for instance, the sequence with all the bits equal to 0. At each time step, we perform the following operations: 1) we randomly extract one of the already existing leaf sequences, say $\bar{s}$, 2) with probability τ, a randomly extracted portion of length $l/4$ of $\bar{s}$ is replaced with the corresponding portion of another randomly chosen sequence (The choice of $l/4$ is arbitrary but does not bring loss of generality. Choosing randomly in the interval $[0, l/4]$, the length of the part of the sequence horizontally transferred does not alter the qualitative behavior of the reconstructing algorithms (results reported in the Supplementary Material online). This last procedure is adopted in the four-state two-parameters Kimura model (see below).), 3) $\bar{s}$ generates two clones as descendants, and 4) each site of the two new sequences is independently flipped with probability m/l, where m is extracted from an exponential distribution with average μ (average number of mutations per sequence per time step). To ensure that at least on site mutates at each branching event, we randomly choose a site to mutate if no site mutated. We iterate this procedure until the desired number of taxa is obtained.

We here consider as distance between two taxa the “correct hamming distance” (Felsenstein 2004), defined as:

$$\mathscr{D}_{\text{corr}} = -\frac{1}{2}\ln(1 - 2h), \quad (5)$$

where h is the “hamming distance,” defined as the fraction of sites in which the sequences differ (In all the results reported in this paper, we let the algorithm infer the correct phylogeny by using the correct hamming distance $\mathscr{D}_{\text{corr}}$. Even though the defined correction has its theoretical justification only in absence of horizontal gene transfer, we checked (data not shown) that using the hamming distance h, all the considered algorithms show the same relative behavior as in the reported results, but the absolute performances are remarkably poorer.).

Although the evolutionary model described above is a toy model for describing evolution, it allows to control and to tune noise as well as horizontal transfer events. We also test our algorithm on phylogenies constructed following more realistic model of evolution, such us the standard four-states two-parameters Kimura model (Kimura 1980). In particular, we follow the same steps described above for the two-state model, but we now consider sequences of nucleotides, with an alphabet of four letters, and different rates of transitions (α) and transversions (2β). We consider in this case as distance between two taxa the correct hamming distance for the Jukes–Cantor model (Felsenstein 2004), which is the limit of the Kimura model when $\alpha = \beta$:

$$d^{\mathrm{JC}}_{\mathrm{corr}} = -\frac{3}{4}\ln\left(1-\frac{4}{3}h\right). \qquad (6)$$

Robinson–Foulds Measure

In order to assess the performances of the different algorithms to reconstruct the true phylogeny, we consider the standard Robinson–Foulds (RF) measure (Robinson and Foulds 1981), which counts the number of bipartitions on which the inferred tree differs from the true one. A bipartition is a split of the leaves in two sets realized through a cut of a tree edge. We recall that it exists a one-to-one correspondence between the bipartitions of the tree and the set of its edges, so that each tree is uniquely characterized by the set of bipartitions it induces.

Competing Algorithms

In order to assess the performances of our algorithm, both in its simulated annealing–like and in its greedy versions, we compare it with the NJ (Saitou and Nei 1987) and the FastME (Desper and Gascuel 2002) algorithms. In the following, we will refer to the simulated annealing–like version of SBiX, unless we will explicitly state the opposite. In addition, we implemented Pauplin's length minimization (from now onward referred as PAUPLIN) by making use of our SBiX algorithm in its form with $k = 0$ (see the above section about the algorithm's description) in order to directly investigate the effectiveness of a nongreedy minimization of Pauplin's length in reconstructing trees. Finally, we implemented the version of our algorithm without the Pauplin's weights (NQ; as discussed above). We also show a comparison with the performances of a state-of-the-art character-based algorithm, MrBayes (Huelsenbeck and Ronquist 2001).

Performances of the Different Algorithms

In this section, we compare the performances of all the considered algorithms as a function of the mutation rate and the horizontal transfer rate in the underlying evolutionary processes described in the Competing Algorithms section.

In figure 3, we show the RF curves for different algorithms (left) as a function of the mutation rate for a fixed tree size ($N = 60$) and $k = 5$. In the whole range of values of the mutation rate, all the versions of our algorithm (PAUPLIN, NQ, and SBiX) outperforms both NJ and FastME. In particular, SBiX outperforms all the other algorithms. Differences between the global minimization of Pauplin's length (PAUPLIN) and FastME arise for very high mutation rates, where the global Pauplin's length minimization outperforms FastME. This is probably due to the fact that FastME is time optimized and therefore less able of our SLS scheme to find the global minimum of the functional for very high mutation rates (for a discussion on the consistency of greedy local moves based on the balanced minimum evolution principle, see Bordewich et al. 2009). In figure 3, we report the dependence of the SBiX performances (right) on the value of the parameter k. It is evident the existence of a range of values between $k = 5$ and $k = 10$ where the algorithm features the best results in a stable way. In the following, unless otherwise stated, we will consider the $k = 5$ case.

Up to this point, we have characterized the performance of the different algorithms for fixed value of the number of leaves, that is, for a given system size. We are now interested in the robustness of our results at different number N of leaves of the tree. Defining $\lambda(N)$ as the mean topological distance between any couple of leaves, we empirically found that each algorithm can be characterized by a reference curve obtained by plotting the normalized RF distance as a function of $\mu^2\lambda(N)$. This scaling can be understood by considering that the relevant quantity for the tree reconstruction is not the bare mutation rate but the amount of back mutation events that can be estimated as $\mu^2\lambda(N)$.

The scaling of the normalized RF distance when reconstructing trees of different sizes is shown in figure 4, where for the sake of clarity, we only report the curves for FastME and SBiX algorithms. Each of the two algorithms is characterized by a different reference curve, and the interesting point here is that the SBiX algorithm is systematically better than FastME at all mutation rates and sizes. We use both the measured value of $\lambda(N)$ in the simulated phylogenies (fig. 4*a*), and the value analytically calculated in the case of perfectly balanced rooted trees (fig. 4*b*), which reads

$$\lambda_{\mathrm{theo}}(N) = \frac{2N(\log_2 N + 1) - 4N + 2}{N - 1}. \qquad (7)$$

We now consider the ability of the different algorithms in recovering the correct tree in presence of horizontal transfer events. The RF curves at a fixed tree size are shown as a function of the probability τ for each sequence to receive a borrowing (with the mechanism defined above). In figure 5*a*, results at low mutation rate are reported, when deviation from additivity is almost exclusively due to the horizontal transfer events. In figure 5*b*, instead, results are reported at high mutation rate, when both back mutations and the horizontal transfer events are responsible for deviations from additivity. For horizontal transfer events cooccurring with a low mutation rate, our algorithm is the most suitable to recover the correct tree, at each rate τ. The NQ method, conversely, shows a performance lower than that of NJ. When a high mutation rate is considered jointly with horizontal transfer events, our SBiX algorithm significantly outperforms the others when the probability τ of horizontal transfer is not too high, whereas in the high τ region, the performance of our algorithm becomes comparable to the minimization of Pauplin's length (PAUPLIN).

It is interesting to compare the performance of the different algorithms in the case of a more realistic data generator. In figure 6, we show the analogous of figure 3 for the two-parameter Kimura model: We present the RF curves obtained for different algorithms (NJ, FastMe, SBiX greedy starting from FastME, SBiX greedy starting from NJ, SBiX simulated annealing–like, and MrBayes) for $N = 60$ as a function of different per-site mutation rates μ/l. The details

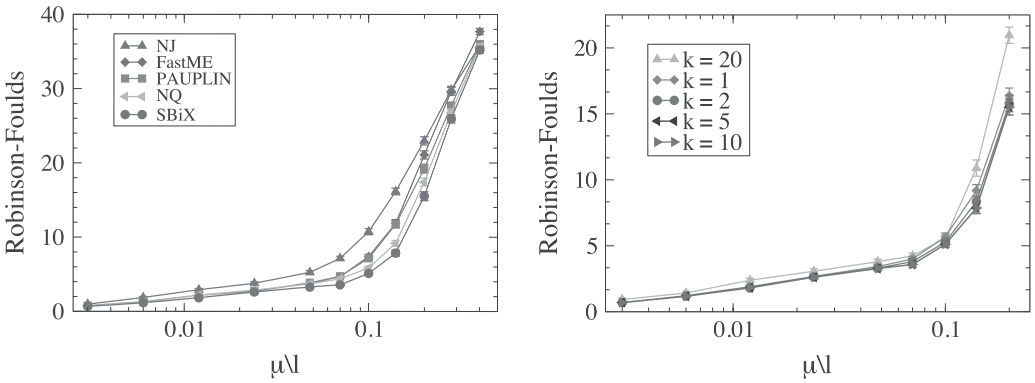

FIG. 3. Performances comparison as a function of the mutation rate. Left: RF distance between the reconstructed and the true trees as a function of the mutation rate per site of the generative evolutionary model. The horizontal transfer rate τ is here kept $\tau = 0$. We compare the performances of the SBiX algorithm with $k = 5$ with: NJ, FastME, the Pauplin length minimization (PAUPLIN), and NQ. Right: dependence of the SBiX algorithm on the parameter k. The best performances are very stable in the range of k between 5 and 10. In both figures, results are averaged over 100 independent realizations for each reported mutation rate. The error bars are standard errors. All the trees generated have $N = 60$ leaves and the sequences have fixed length $l = 1,000$.

of the simulation for MrBayes are presented in the Supplementary Materials online together with a comment on its computational complexity.

The first evidence is that SBiX, in its different versions, clearly outperforms the other two distance-based algorithms (NJ and FastME). The improvement is even more evident than in the binary characters case displayed in figure 3. The three variants of SBiX perform similarly in the low mutation rate regime ($\mu/l < 0.2$), and the results show a moderate improvement of the simulated annealing–like version only for high level of mutation rate, whereas the difference between the two greedy versions of SBiX (the one using as a starting point the tree reconstructed by the NJ algorithm and the other by the FastME algorithm) seems to be statistically irrelevant in all the mutation rate interval analyzed.

Let us now discuss the comparison with MrBayes. After a very low mutation rate regime ($\mu/l \leqslant 0.024$), where all algorithms show analogous accuracies, one enters a regime where MrBayes outperforms SBiX. We should remark that the precise estimate of the MrBayes performance, especially, in the high mutation regime, is problematic due to the the issue of reaching a steady state for a Monte Carlo Markov Chain. For assessing the convergence of MrBayes, we monitored (see Supplementary Materials online for details) both the partition variance and the posterior likelihood time series, and, especially for the larger mutation rate, a fraction of the samples seems indeed not to have reached convergence. One has to note that in this regime, as more thoroughly discussed in the Supplementary Materials online, the computational time for a single sample is already of roughly 7 h, whereas for both greedy versions of SBiX is of the order of 10^{-2} s and for the simulated annealing–like version is around 5 min.

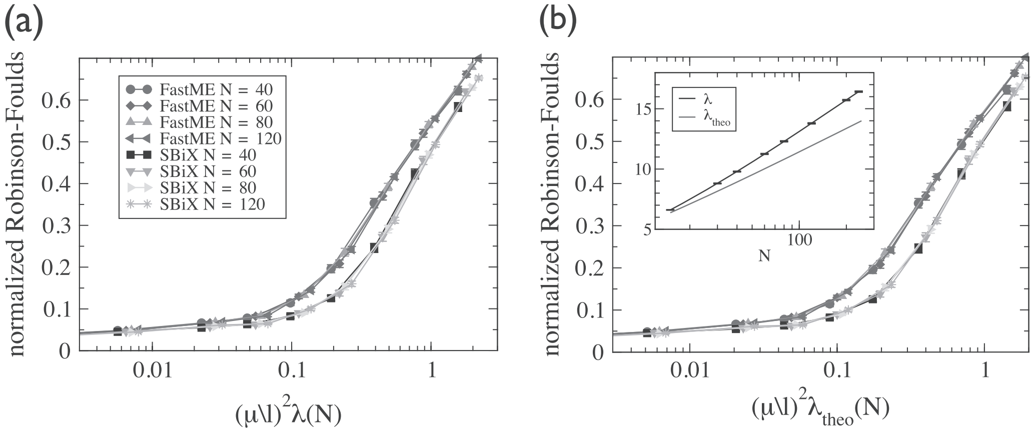

FIG. 4. System-size dependence. Behavior of the normalized RF distance for the SBiX algorithm and FastME for different system sizes, that is, different values N of the number of leaves. Here, normalized means the RF distance divided by its maximal value $N - 3$. In all the cases, curves for different values of N collapse as a function of $\mu^2\lambda(N)$ (see text for details), where $\lambda(N)$ is the average distance between two leaves in a tree with N leaves. In both the analysis, the horizontal transfer rate τ is kept $\tau = 0$. We use both the true value of $\lambda(N)$ in the simulated phylogenies (a) and the value analytically calculated in the case of perfectly balanced trees $\lambda_{\text{theo}}(N)$ (b). In the inset of (b), we report the behavior of λ and λ_{theo} as a function of tree size N. The experimental values $\lambda(N)$ are systematically larger than λ_{theo} putting in evidence a slight deviation of the generated trees from a perfectly balance condition.

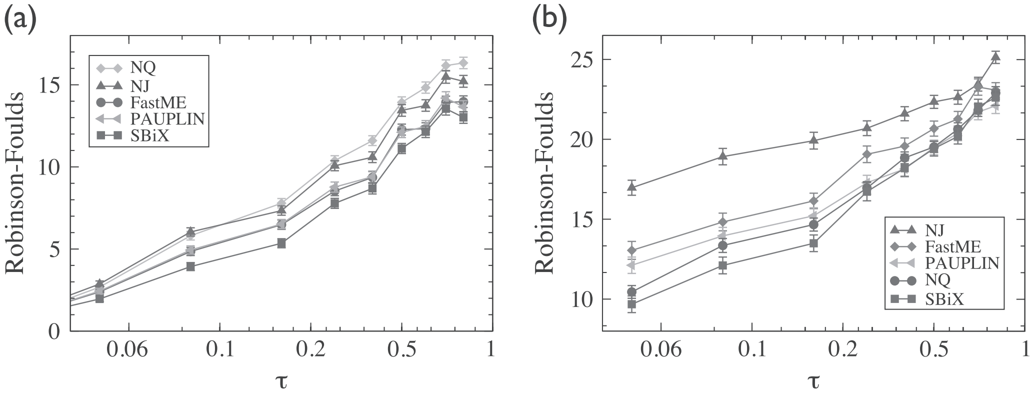

FIG. 5. Performances comparison as a function of the horizontal transfer rate. RF measure for the reconstructed trees as a function of the horizontal transfer rate of the generative model. We compare the performances of our SBiX algorithm (with $k = 5$) with those of the following ones: NJ, FastME, Pauplin's length minimization algorithm (PAUPLIN), and the NQ algorithm. Results are averages over 100 independent realizations for each reported horizontal transfer rate. The error bars are standard errors. All the trees generated have $N = 60$ leaves. (*a*): mutation rate per site $\mu/l = 0.03$, whereas the sequences have fixed length $l = 10,000$. (*b*): mutation rate per site of $\mu/l = 0.14$, whereas the sequences have fixed length $l = 1,000$.

Discussion and Conclusions

In this paper, we have introduced a new algorithmic scheme for phylogeny reconstruction. Belonging to the family of SLS algorithms, our scheme crucially exploits two-known properties of additive distance matrices, the four-points condition and the so-called Pauplin's length. We proposed in particular a stochastic scheme where the correct topology is inferred through a series of swapping of the tree topology. When tested on artificially generated phylogenies, our algorithmic scheme significantly outperforms state-of-art distance-based algorithms in cases of deviation from additivity due to high rate of back mutations. A significant improvement is also observed with respect to the state-of-art algorithms in case of high rate of horizontal transfer.

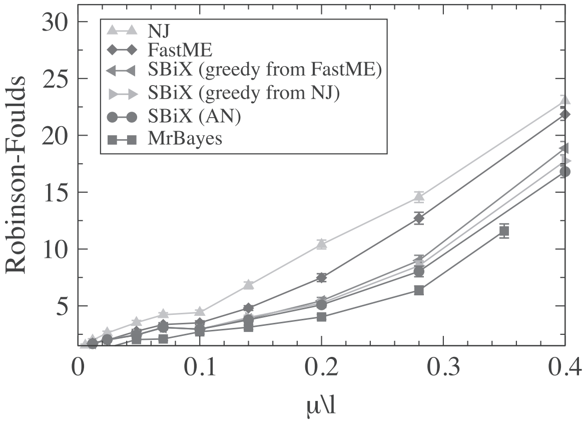

FIG. 6. Performances comparison as a function of the per-site mutations rate. The evolutionary model used to generate phylogenies is the Kimura two-parameters model with transition rate $\alpha = 0.4$ and transversion rate $2\beta = 0.6$. RF measure for the reconstructed trees as a function of the mutation rate per site. We compare the performances of the following algorithms: NJ, FastME, SBiX greedy from NJ, SBiX greedy from FastME, SBiX simulated annealing, and MrBayes. We set $k = 5$ in both the greedy and the simulated-annealing versions of SBiX. Results are averages over 100 independent realizations for each reported mutation rate. The error bars are standard errors. All the trees generated have $N = 60$ leaves and the sequence $l = 1,000$.

Such good performances are due to the way we differentially weight the different quartets contributions with a term inversely proportional to their length and thus to their probability to be affected by back mutations. On the other hand, further work is needed for a complete theoretical understanding of the algorithm. In particular, despite many attempts, we are, at present, unable to formulate the update strategy in terms of a state functional. Beside the interest in itself, this would open the way to analytic treatments as well as to algorithmic optimization strategies possibly more efficient than the SLS one.

As for the comparison of our algorithmic scheme with state-of-the-art algorithms, it is fair to observe that SBiX features a definitely larger computational complexity but, in practice, its greedy version is fast enough for reconstructing phylogenies up to a few thousands of leaves.

Though SBiX outperforms all competitors also in presence of horizontal transfers, the method is especially suited for dealing with nonadditivity originated by back mutations. The issue of horizontal transfer is, however, central in many fields (Doolittle et al. 2008), and we believe that formulating effective strategies for dealing with it, considering both phylogenetic trees and networks, is an open challenge for the next generation reconstruction algorithms and will be the aim of further studies.

It is worth mentioning how the applicability of phylogenetic algorithms has recently widened its scope. Many different fields have arisen in the last few years where a correct reconstruction of phylogenetic trees may reveal underlying relevant dynamical processes. For instance, phylodynamics is a new field at the crossroad of immunodynamics, epidemiology, and evolutionary biology, which explores the diversity of epidemiological and phylogenetic patterns observed in RNA viruses of vertebrates (Grenfell et al. 2004); phylogeography is the study of the historical processes that may be responsible for the contemporary geographic distributions of individuals as well as of languages or viruses (Avise 2000). In all these cases, a strong effort is being devoted to the collection of comprehensive data sets, and efficient and reliable algorithms are needed especially when deviations from perfect phylogenies become relevant.

Supplementary Material

Supplementary Material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank Simone Pompei, Francesca Colaiori, Ciro Cattuto, and Damián H. Zanette for many interesting discussions and suggestions. This research has been partly supported by the TAGora and ATACD projects funded by the European Commission under the contracts IST-34721 and 043415.

Appendix A: The Equivalence with Pauplin's Length Minimization

We show here that our SBiX algorithm minimizes Pauplin's length when $k = 0$ in the equation (2). In order to see this, we explicitly calculate the cost difference between, say, the configurations $((A,B),(C,D))$ and $((A,C),(B,D))$. We note that the sum on the considered quartets can be divided in three parts in which one of the three distances D_1, D_2, and D_3 is, respectively, minimal (where, as already defined in the text, $D_1 = \mathscr{D}(a,b) + \mathscr{D}(c,d)$, $D_2 = \mathscr{D}(a,c) + \mathscr{D}(b,d)$, and $D_3 = \mathscr{D}(a,d) + \mathscr{D}(b,c)$). After a little algebra, one gets

$$\Delta E = \sum (D_2 - D_1) 2^{-t(a,\alpha)-t(b,\beta)-t(c,\gamma)-t(d,\delta)}, \qquad (8)$$

where the sum is again over the $a \in A, b \in B, c \in C, d \in D$ and $\Delta E \equiv E_{((A,C),(B,D))} - E_{((A,B),(C,D))}$. Making use of the relation

$$\sum_i 2^{-t(i,r)} = 1, \qquad (9)$$

and of the equivalences: $2^{-t(a,b)} = 2^{-t(a,\alpha)-t(b,\beta)}/2$ in the configuration $((A,B),(C,D))$ and $2^{-t(a,b)} = 2^{-t(a,\alpha)-t(b,\beta)}/4$ in the configuration $((A,C),(B,D))$ (and the analogous relations for the other pairs of taxa), it is easy to prove that it holds

$$\Delta E = 4 \Delta L_{\mathrm{P}}, \qquad (10)$$

where L_{P} is Pauplin's length and $\Delta L_{\mathrm{P}} \equiv L_{\mathrm{P},((A,C),(B,D))} - L_{\mathrm{P},((A,B),(C,D))}$ is the difference of the Pauplin's length between the two configurations.

Appendix B: Locality of the SBiX Configuration Cost

We give here an argument to prove that differences in the local cost of our SBiX method cannot be written as differences of a functional on the whole tree. If this was the case, a functional could be defined as $F(x) = F(x_0) + \sum \Delta E_i$, where $F(x_0)$ is the value taken by the functional in a reference configuration x_0, and ΔE_i are the cost differences along a path from x_0 to x. Moving in the space of tree's topologies, we should obtain the same value of F each time we visit the same topology, that is, the difference of cost between two states does not depend on the path. This is not the case, as we explicitly checked, when the cost is defined as in equation (3) and $k \neq 0$.

References

Atteson K. 1997. The performance of neighbor-joining algorithms of phylogeny reconstruction. In: Jiang T, Lee D, editors. Lecture Notes in *Computer Science*. 1276. Berlin, Germany:Springer-Verlag. p. 101–110.

Avise JC. 2000. Phylogeography: the history and formation of species. Cambridge (MA): Harvard University Press.

Bordewich M, Gascuel O, Huber KT, Moulton V. 2009. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Trans Comput Biol Bioinform.* 6(1):110–117.

Bruno WJ, Socci ND, Alpern AL. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol.* 17:189–197.

Buneman P. 1971. The recovery of trees from measures of dissimilarity. In: Hudson FR, Kendall DG, Tautu P, editors. *Mathematics in archeological and historical sciences*. Edinburgh (UK): Edinburgh University Press. p. 387–395.

Buneman P. 1974. A note on the metric properties of trees. *J Comb. Theory Ser B.* 17:48–50.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet.* 19:233–257.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol.* 9:687–705.

Doolittle WF, Nesbo CL, Bapteste E, Zhaxybayeva O. 2008. Lateral gene transfer. In: Pagel M, Pomiankowski A, editors. *Evolutionary genomics and proteomics*. Sunderland (MA): Sinauer. p. 45–79.

Erdös PL, Rice K, Szekely LA, Warnow TJ, Steel M, Warnow YJ. 1998. The short quartet method. In: Proceedings of International Congress on Automata, Languages and Programming, ICALP'98. July 13–17, 1998. Aalborg, Denmark. Springer. Lecture Notes in Computer Science.

Felsenstein J. 2004. *Inferring phylogenies*. Sunderland (MA): Sinauer Associates Inc.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.

Gascuel O, editor. 2007. Mathematics of evolution and phylogeny. Oxford (UK): Oxford University Press.

Gascuel O, Steel M. 2006. Neighbor-joining revealed. *Mol Biol Evol.* 23:1997–2000.

Gray RD, Atkinson QD. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303:327–332.

Gusfield D. 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. New York: Cambridge University Press.

Hoos HH, Stützle T, 2005. Stochastic Local Search: Foundations and Application. Amsterdam, The Netherlands: Morgan Kaufmann, Elsevier.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformmatics* 17(8):754–755.

Kimura M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.

Kirkpatrick SJ, Gelatt CD, Vecchi MP. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Lazer D, Pentland A, Adamic L, et al. (15 co-authors). 2009. Social science: computational social science. *Science* 323:721–723.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324:1561–1564.

Mihaescu R, Levy D, Pachter L. 2007. Why neighbor-joining works. *Algorithmica* 54:1–24.

Pauplin Y. 2000. Direct calculation of a tree length using a distance matrix. *J Mol Evol.* 51:41–47.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet.* 10:540–550.

Ragoussis J. 2009. Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet.* 10:117–133.

Renfrew C, McMahon A, Trask L. 2000. Time depth in historical linguistics. Cambridge (UK): The McDonald Institute for Archeological Research.

Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Simonson AB, Servin JA, Skophammer RG, Herbold CW, Rivera MC, Lake JA. 2005. Decoding the genomic tree of life. *Proc Natl Acad Sci U S A.* 102:6608–6613.

Snir S, Warnow T, Rao S. 2008. Short quartet puzzling: a new quartet-based phylogeny reconstruction algorithm. *J Comput Biol.* 15: 91–103.