BMC
Evolutionary Biology

METHODOLOGY ARTICLE

Open Access

EM for phylogenetic topology reconstruction on nonhomogeneous data

Esther Ibáñez-Marcelo[1*] and Marta Casanellas[2]

Abstract

Background: The reconstruction of the phylogenetic tree topology of four taxa is, still nowadays, one of the main challenges in phylogenetics. Its difficulties lie in considering not too restrictive evolutionary models, and correctly dealing with the long-branch attraction problem. The correct reconstruction of 4-taxon trees is crucial for making quartet-based methods work and being able to recover large phylogenies.

Methods: We adapt the well known expectation-maximization algorithm to evolutionary Markov models on phylogenetic 4-taxon trees. We then use this algorithm to estimate the substitution parameters, compute the corresponding likelihood, and to infer the most likely quartet.

Results: In this paper we consider an expectation-maximization method for maximizing the likelihood of (time nonhomogeneous) evolutionary Markov models on trees. We study its success on reconstructing 4-taxon topologies and its performance as input method in quartet-based phylogenetic reconstruction methods such as QFIT and QuartetSuite. Our results show that the method proposed here outperforms neighbor-joining and the usual (time-homogeneous continuous-time) maximum likelihood methods on 4-leaved trees with among-lineage instantaneous rate heterogeneity, and perform similarly to usual continuous-time maximum-likelihood when data satisfies the assumptions of both methods.

Conclusions: The method presented in this paper is well suited for reconstructing the topology of any number of taxa via quartet-based methods and is highly accurate, specially regarding largely divergent trees and time nonhomogeneous data.

Keywords: Tree topology reconstruction, Expectation-maximization, Quartet-based method, Evolutionary Markov model

Background

Obtaining a good method for reconstructing the phylogenetic topology of four taxa is one of the crucial goals in phylogenetics. The four-taxon trees, if correctly inferred, can be used as input of quartet-based methods in order to reconstruct larger trees. But due to the complexity of real data, the problem of reconstructing four-taxon trees is not so easy. Most phylogenetic reconstruction methods assume simple evolutionary models that may not really fit real data, which leads to incorrect phylogenetic inference ([1-5]). For example, many of them

rely on continuous-time Markov processes with a constant instantaneous mutation rate matrix along the tree (the so-called global *homogeneity*), and also assume time-reversibility (and hence stationarity). On the other hand, as pointed out in [6], in order to make quartet-based methods work it is extremely important to obtain 4-taxon tree reconstruction methods that are not affected by the presence of long-branch attraction [7].

Most evolutionary models are described by a Markov process over the tree, that is, conditional rates of change at two diverging sequences depend only on the current state and are independent of the previous sates ([8], chapter 8.2). *Markov processes* on trees are specified by a distribution at the root of the tree and a transition matrix at each branch and, in contrast to continuous-time models, the Markov process on each branch is not assumed to be time-homogeneous [9] (that is, they are *locally* heterogeneous).

*Correspondence: eibanez@crm.cat
[1]Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C - 08193 Bellaterra, Barcelona, Spain
Full list of author information is available at the end of the article

These models directly consider as parameters the entries of the substitution matrices and the root distribution (see [8], chapter 8, [10], § 4.2). Barry and Hartigan [11] considered such a general Markov model (henceforth called GMM), which does not assume any other constraints. In particular, it is locally and *globally* time nonhomogeneous (instantaneous substitution rates are not constant on any edge, nor on the whole tree) and it is neither time-reversible nor stationary. The only restriction underlying this model is that sites evolve independently and are identically distributed (i.i.d. hypothesis). Considering this model and its submodels is one way of covering more general scenarios, in contrast to those phylogenetic methods that implement time-homogeneous and time-reversible continuous-time models (GTR and its submodels) cf [9].

The GMM above accounts for 12 parameters per edge plus three parameters for the root distribution. When some symmetries on the transition matrices or on the root distribution are imposed, one obtains the substitution matrices of the corresponding Jukes-Cantor and Kimura (2 and 3 parameters) models among others (([10], § 4.2),[12], see Methods section). For instance, the Markov version of the K81 model [13] (henceforth referred to as K81*) deals with 3 parameters per edge (one for the conditional probability of transitions, and two for the two types of transversions, see the Methods section) which makes a total of $3 * (2n - 3)$ parameters in unrooted trivalent trees with n leaves, whereas the usual time-homogeneous continuous-time version accounts for 2 parameters for a normalized instantaneous rate matrix constant over the tree plus one parameter per edge length (that is, $2 + (2n - 3)$ parameters on an unrooted trivalent tree). Notice, however, that if one considers a time nonhomogeneous continuous-time Kimura 3-parameter model, then the number of parameters is exactly the same as for K81*. In this case, the only difference between both models is that K81* does not even assume local homogeneity (that is, time homogeneity over each branch), while all time-continuous models do. The huge amount of parameters for nonhomogeneous models makes a maximum-likelihood approach unfeasible and unreliable for a whole tree on n taxa if n is large. Nevertheless, there is some hope that these more general models lead to accurate methods on 4-taxon trees. In our setting, we only deal with substitutions of nucleotides (not aminoacids) on 4-taxon trees and we will always assume the i.i.d. hypothesis (thus excluding the possibility of heterogeneity across sites).

In this paper we develop a maximum likelihood framework for inferring the best tree topology under (general) Markov processes. Our approach is based on the widely used Expectation-Maximization algorithm. The Expectation-Maximization algorithm (EM), as introduced in [14], is an iterative algorithm for finding maximum likelihood estimates of parameters in statistical models that deal with unobserved data. We have adapted this algorithm to the case of phylogenetic 4-taxon trees in what we call `EMtree`. EM iteratively gives an expectation of the distribution of the nucleotide sequences at the interior nodes (this is called the E-step) and finds the parameters that maximize the likelihood for these data in the so-called M-step (because the parameters for which the maximum likelihood is achieved can be computed in a closed form for complete data). The EM algorithm has been applied to many other disciplines (see for example [15]).

The use of the EM algorithm to estimate the continuous parameters of a phylogenetic tree under a Markov process (namely, the root distribution and the entries of the transition matrices) has been already discussed in [9] and [16]. In this paper we focus on the use of `EMtree` for estimating the tree topology for four taxa and test its performance in reconstructing the correct tree topology on simulated data. We shall see that, although it is a time consuming algorithm, it can lead to very accurate results.

First of all we test it as a 4-taxon tree reconstruction method by using the tree space proposed by Huelsenbeck [17] on data simulated from a general Markov process first and then restricting to a time-homogeneous (continuous-time) process. We compare `EMtree` to neighbor-joining and to the usual (continuous-time) maximum likelihood approach under both global homogeneity and nonhomogeneity (note that throughout the experiments we will restrict to time-reversible models to simulate and to recover trees). As all models considered will be stationary (and with uniform stationary distribution), we will not be evaluating compositional heterogeneity but only the effect of the variation of substitution rates among lineages [18]. Afterwards, we use these three methods as input of two quartet-based methods: one weighted (Willson's Quartet-Suite [19,20]) and one unweighted (Maximum Quartet Fit, QFIT as implemented in Clann [21]); and compare their performance on the 12-taxon trees proposed in [6]. Specially for largely divergent trees, we observe that `EMtree` gives the best results and is less subject to long-branch attraction.

Results and discussion

In this section we present the results of the topology reconstruction method `EMtree` on simulated data evolving both under homogeneous and nonhomogeneous Markov processes.

In Figure 1 we present the performance of three different topology reconstruction methods on the 4-taxon tree space proposed by Huelsenbeck [17] (see Methods section): Figure 1(a) for the usual neighbor-joining (`NJ`) with the K81 distance, 1(b) for the usual maximum likelihood tree (`ML`) assuming time-homogeneous continuous-time K81 model, and 1(c) `EMtree` for K81* model. The Huelsenbeck tree space covers a wide range of branch
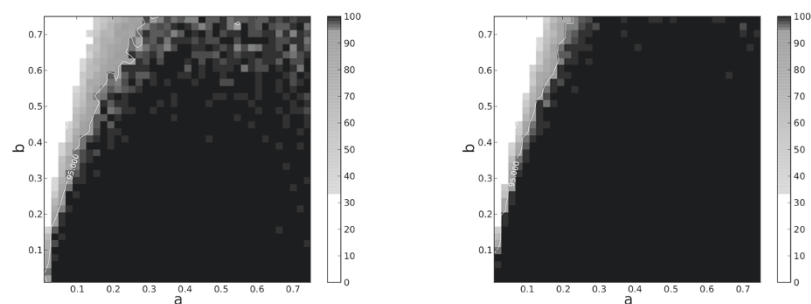
(a) Results corresponding to NJ on alignments of length 300 bp. (mean: 90.76, sd: 24.91) *(left)* and 1000 bp. (mean: 91.45, sd: 25.73) *(right)*.
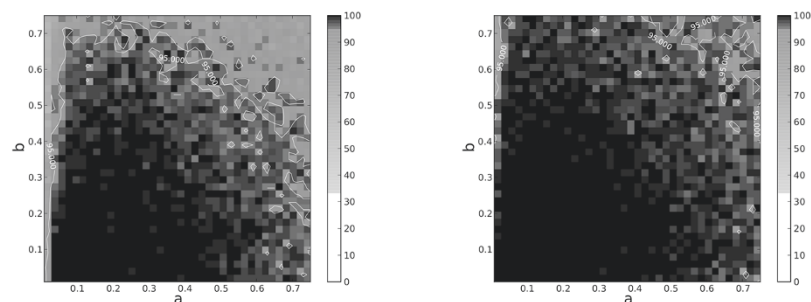
(b) Results corresponding to time-homogeneous ML on alignments of length 300 bp. (mean: 96.52, sd: 4.44) *(left)* and 1000 bp. (mean: 98.51, sd: 2.05) *(right)*.
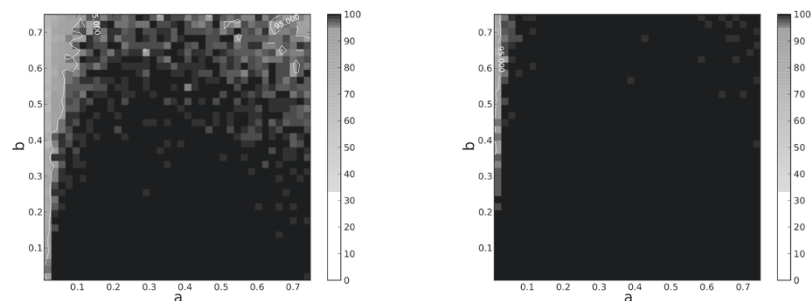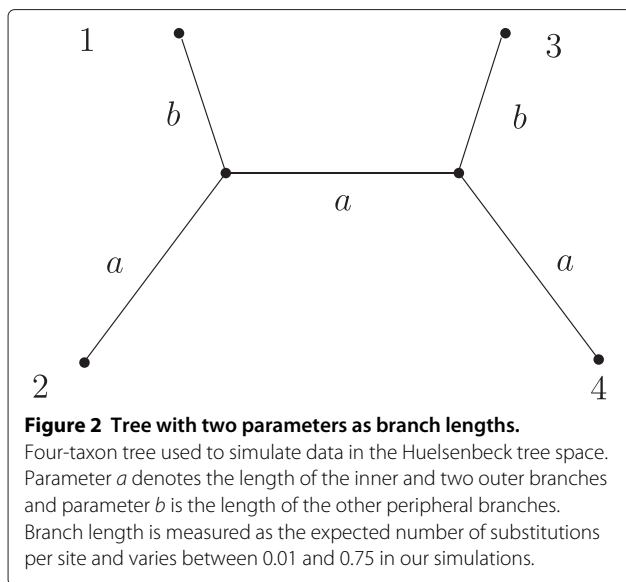
(c) Results corresponding to EMtree on alignments of length 300 bp. (mean: 98.30, sd: 4.42) *(left)* and 1000 bp. (mean: 99.79, sd: 1.47) *(right)*.

Figure 1 Results of three different topology reconstruction methods: NJ, ML and EMtree. Results of the three different topology reconstruction methods NJ **(a)**, ML **(b)** and EMtree **(c)** on the tree space proposed by Huelsenbeck on K81* data (x-axis corresponds to parameter *a* in Figure 2 and y-axis to parameter *b* in the same figure). Black cell is drawn if the topology is 100% correctly reconstructed on the simulated data. White square denotes 0 to 33% success, and the gray scale in between. Left figures correspond to alignments of lengths 300 bp and right to 1000 bp.

lengths for 4-taxon trees: parameters *a* and *b* denote branch lengths as indicated in Figure 2 and they were varied from 0.01 to 0.75. For each pair (*a*, *b*) we draw a black cell if the topology is correctly reconstructed on the 100 simulated alignments; a white cell denotes less than 33% success; and the gray scale used in between is shown in the figures. The alignments were generated following the K81* model (see the Methods section) and had length 300 (Figure 1 left) and 1000 bp (Figure 1 right).

The results show that NJ has difficulties in the Felsenstein zone where the long-branch attraction problem is present (that is, small *a* and large *b*) and ML fails for largely divergent trees (note that trees here have at most 2.25 pairwise divergence, whereas similar studies have even used divergences of about 7.0 expected substitutions per site [18]). The EMtree algorithm seems to overcome these two difficulties and is clearly more accurate than NJ and ML in this tree space. Nevertheless, we need to point out that the data were simulated according to K81* model, which fits the assumptions of EMtree but not of NJ nor ML methods (indeed, these last two methods are based on the continuous-time Kimura model and

Figure 2 Tree with two parameters as branch lengths.
Four-taxon tree used to simulate data in the Huelsenbeck tree space.
Parameter *a* denotes the length of the inner and two outer branches
and parameter *b* is the length of the other peripheral branches.
Branch length is measured as the expected number of substitutions
per site and varies between 0.01 and 0.75 in our simulations.

assume homogeneous mutation rates along lineages). This model misspecification for ML leads to in incorrect inference of parameters, which is even more extreme for long branches [18] (this justifies the bad performance of ML in the lower right corner of Figure 1(b)). To confront this situation, we also generated data under time-homogeneous continuous-time K81 model. The results are depicted in the left column of Figure 3 and clearly show a similar success of ML and EMtree in this case.

In order to compare EMtree to time nonhomogeneous but continuous-time ML, we generated data evolving under a continuous-time nonhomogeneous K81 model. This is a special case of K81* and therefore this model matches both approaches: EMtree (on its K81* version) and bppml [22] restricted to a nonhomogeneous continuous-time K81 model. On Figure 3 (right) we present the results of these two methods on nonhomogeneous K81 data to show that both methods perform similarly and outperform NJ.

Although EMtree may give more accurate results on more general data, it is a time consuming algorithm. Indeed, on 4-taxon trees, EMtree is almost 1000 times slower than NJ and more than 2000 times slower than ML. For example on 100 4-taxon alignments of 600bp, the execution time on a Intel(R) Core(TM) i5-4200U CPU @ 1.60GHz (only using one of the four CPU) was 186.66s for EMtree on K81*, 0.06s for PAML ML on homogeneous continuous-time K81, 211.45 for bppml on nonhomogeneous K81 model, and 0.16s for NJ (with K81 distances).

Now we present the results that test the use of EMtree, NJ, and (time-homogeneous continuous-time) ML as methods to obtain the input for the quartet-based methods QuartetSuite and QFIT on data generated under K81* model.

For the three topologies on twelve taxa studied here (called *cc, cd*, and *dd* –see Figure 4 and the Methods section), we give the proportion of the (one thousand) reconstructed trees whose Robinson-Foulds distance to the original topology is equal to $0, 2, 4, 6$ or > 6, for both QuartetSuite and QFIT, and for the three input methods under study. The results are displayed in Figures 5 (QuartetSuite) and 6 (QFIT) for simulated alignments of 600 bp and in Figures A1 and A2 of the Additional file 1 for 300 bp. The figures for both alignment lengths are similar and present the same trends (slightly better for 600 bp for all methods, as expected), so we let the results on 300 bp for the Additional file 1. In Figure 5 we also present the performance of a global NJ and a global ML (estimating time-homogeneous continuous-time K81 model).

In both quartet-based methods (QuartetSuite and QFIT), the reconstruction of the *cc* topology presents the best results compared to the others, independently of the algorithm used as input quartets (EMtree, NJ, or ML). The same tendency is shared by global ML and NJ. Conversely, the topology *dd* is never correctly reconstructed for any of the methods or branch lengths. It is worth noting that *cc* is the topology that would have the least long branch attraction and *dd* is the one that would have the most (*cd* is in between because half of the tree comes from *cc* and the other half comes from *dd*). Therefore our results are consistent with the observation in [6] that the success of a quartet based method depends on the capacity of the input method to correctly reconstruct 4-taxon trees under the long-branch attraction problem.

In both Figures 5 and 6 we observe that the performance of ML and EMtree is quite similar in most cases, although ML never outperforms EMtree. A detailed look at these figures reveals that for largely divergent trees (that is, $b = 0.1$, the last bar in each plot), EMtree is the best quartet input method among those considered here, as its results outperform NJ and ML for both QuartetSuite and QFIT and in all tree topologies. We find the explanation of this result in the management of long-branch attraction by the different methods considered. Long branches lead to similar sequences as a result of multiple substitutions and, as ML estimates have been computed on a wrong substitution model, this method is more influenced by long-branch attraction. EMtree has also a better success than a global NJ and a global ML on the trees *cd* and *dd*, but on the "easiest" tree *cc*, a global ML is more accurate. For the largely divergent *cc* tree, a global ML has better average performance than EMtree (see also below), but it never succeeds in fully reconstructing the topology (whereas EMtree does in about 2% of the alignments).

When considering the QuartetSuite method, NJ is clearly the worst quartet input method for all tree topologies and branch lengths. This is probably due to the fact that the Willson method implemented in QuartetSuite
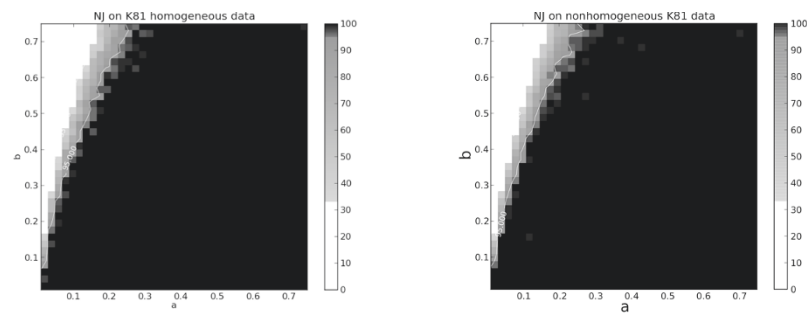
(a) Results corresponding to NJ on 600bp alignments. Data: *(left)* time homeogeneous continuous-time K81 (mean: 90.55, sd: 26.95), *(right)* time nonhomogeneous K81 (mean: 90.62, sd: 26.81).
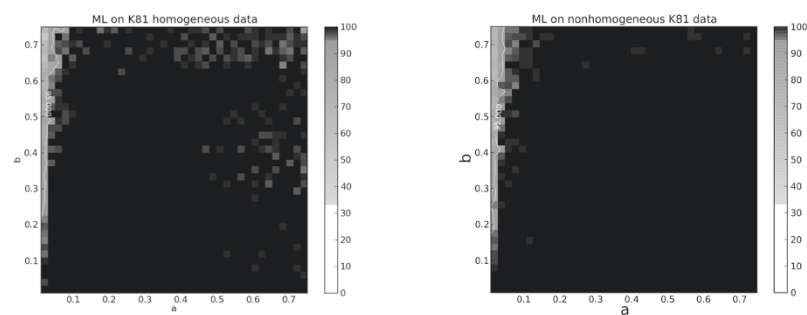
(b) Results corresponding to ML on 600bp alignments. Data: *(left)* time homeogeneous continuous-time K81 (mean: 99.33, sd: 3.22), *(right)* time nonhomogeneous K81 (mean: 99.44, sd: 3.14).
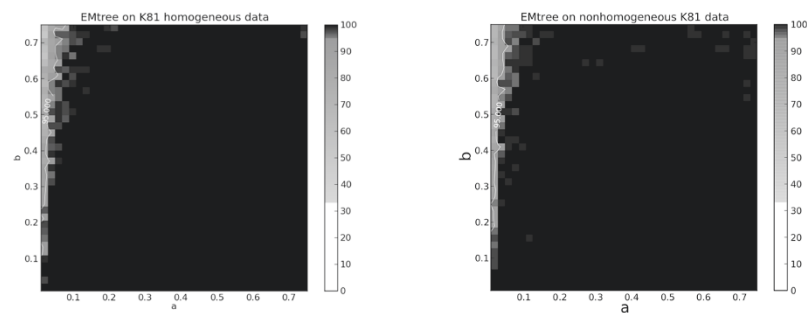
(c) Results corresponding to EMtree on 600bp alignments. Data: *(left)* time homeogeneous continuous-time K81 (mean: 99.35, sd: 3.68), *(right)* time nonhomogeneous K81 (mean: 99.44, sd: 3.11).

Figure 3 Results of three different topology reconstruction methods: NJ, ML and EMtree. Results of the three different topology reconstruction methods NJ **(a)**, ML **(b)** and EMtree **(c)** on the tree space proposed by Huelsenbeck (see Figure 2). On the left, data have been generated using a time-homogeneous continuous-time K81 model and the topology is estimated by NJ (with K81 distance), ML (estimating a homogeneous K81 model), and EMtree (on K81*). On the right, data have been generated using time nonhomogeneous continuous-time K81 model and the topology is estimated by NJ (with K81 distance), ML (estimating time nonhomogeneous K81 continuous-time model), and EMtree (on K81*). The alignments considered in this figure have length 600 bp.

is intended for weighted quartets, whereas NJ quartets are given only binary weights. On the contrary, for the unweighted method QFIT, NJ seems to be more accurate than ML and EMtree for topologies *cc* and *dd* but only for low divergence ($b = 0.005$ or $b = 0.015$).

We want to point out that, in general, QFIT gives worse results than QuartetSuite (except for the NJ algorithm used as input in the *cc* and *cd* topologies with low divergence), reinforcing the idea that weighted methods are more reliable [6,23].

In Tables 1 and 2 we display the mean of the Robinson-Foulds distance of the same study on the 12-taxon trees, and its variance in parentheses (Table 1 for Quartet-Suite, and Table 2 for QFIT). Results are presented for
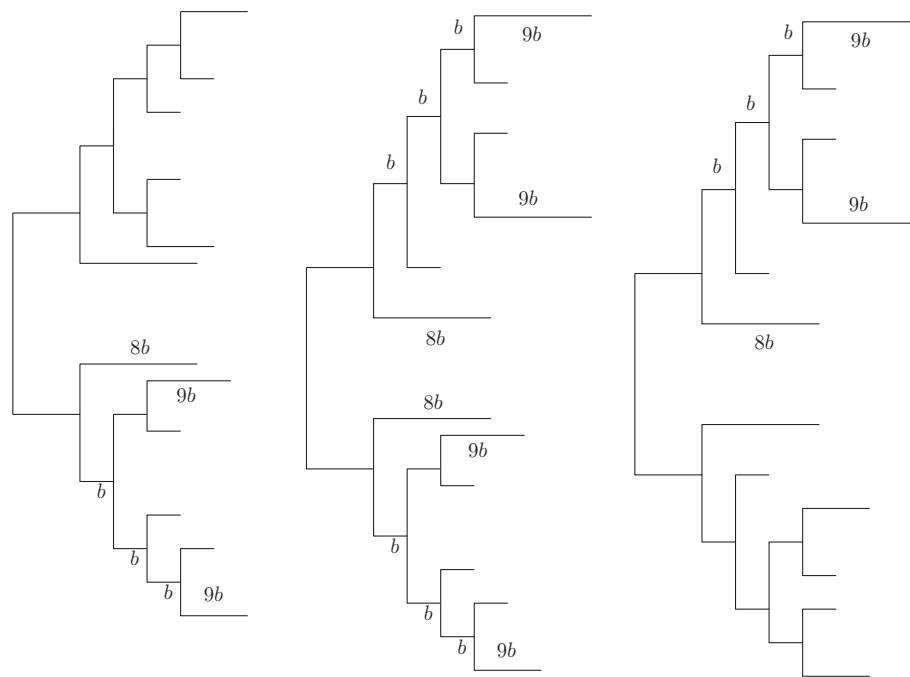
Figure 4 Twelve-taxon topologies used in the paper: *cc, cd ,dd*. Twelve taxa topologies used in the paper, named according to [6]: *cc* (left), *cd* (middle), *dd* (right); *b* is the branch length parameter that varies in {0.005, 0.015, 0.05, 0.1} along the paper.

alignments of 600 bp and have to be interpreted as less mean, better approximation (results for 300 bp are similar and appear in the Additional file 1, Tables S1 and S2). For each tree topology and each choice of parameter *b*, the best method (according to the lowest mean) is marked in bold print.

In these tables we observe that QuartetSuite gives the lowest distance to the original tree in general and that the best results for largely divergent trees ($b = 0.1$) are obtained by EMtree (for both QuartetSuite and QFIT and all tree topologies). Whenever the mean of Robinson-Foulds distance for (quartet-based) ML is lower than the mean for EMtree, there is no significant difference. Overall, EMtree is the method that outperforms the other two quartet input methods in most cases. As far as global methods are concerned, it is worth pointing out the bad performance of a global NJ on trees *cd* and *dd* (or *cc* with $b = 0.05$ or 0.1), specially if one takes into account that the Robinson-Foulds distance for the (less resolved) star tree is 9.
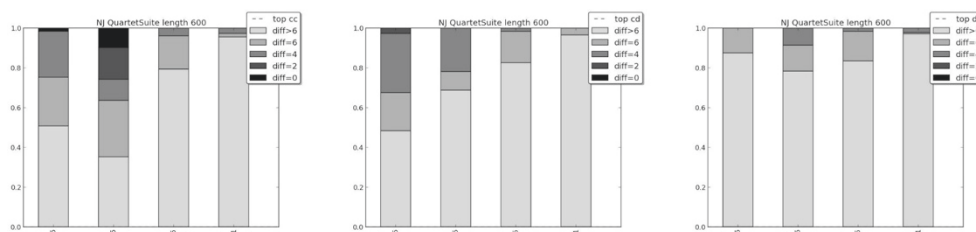
When inspecting the variances, one sees that EMtree is the only quartet input method that preserves low variances in all cases (global methods are the ones presenting lower variance, though). Conversely, the variances are extremely large for NJ with QuartetSuite in all different scenarios, and they are also huge for (quartet input) ML when the trees are largely divergent ($b = 0.1$). QFIT presents low variances in all cases, for all input methods, probably because the input information is less "subject to vary" (the input only considers tree topologies, not weights). It is worth pointing out that whenever quartet input NJ outperforms ML and EMtree (only for QFIT), it does so with larger variance than the other two methods.
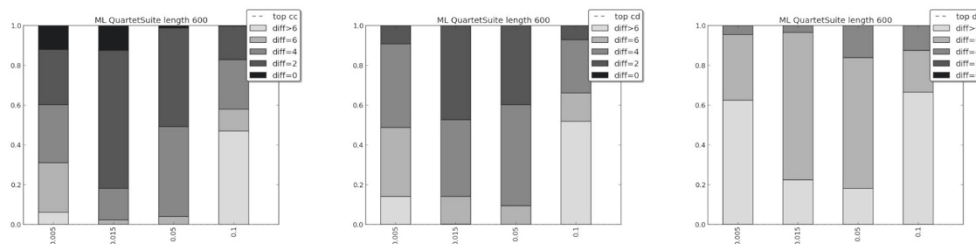
Conclusion
We tested the accuracy of the (likelihood-based) method EMtree as a method to infer 4-taxon topologies under (time nonhomogeneous) Markov models, and compared it to NJ and ML (homogeneous and nonhomogeneous). When EMtree and ML are tested on data satisfying the assumptions of both methods, they have a similar performance. Nevertheless, EMtree is based on time nonhomogeneous models (both local and global time heterogeneity), and hence outperforms the other methods when these assume homogeneity.
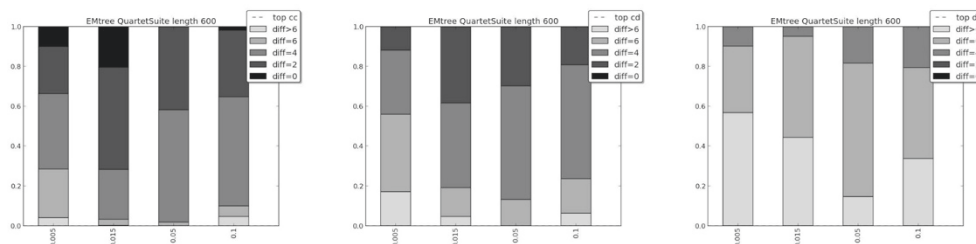
There are only few nonhomogeneous continuous-time models that could be fairly compared to general Markov processes, and one expects that under more complex evolutionary scenarios (such as non-stationary or not time-reversible data), the success of usual ML or NJ methods (based on these assumptions) will be poorer (as shown in [18]), confirming that an EM approach based on general Markov processes, could be more recommendable.
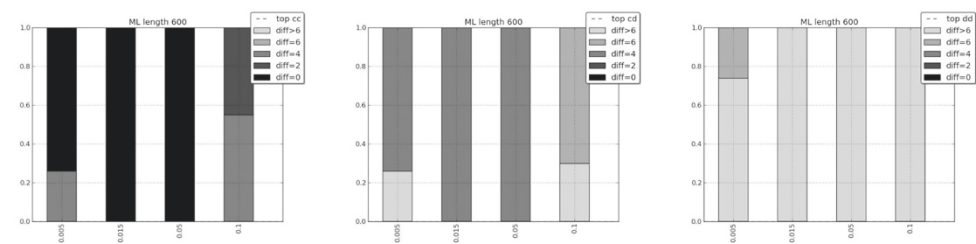
Figure 5 Results of QuartetSuite (quartet-based method) with alignment length equal to 600. Results of QuartetSuite (quartet-based method) for different topologies and different input methods on alignments of length 600. The first three rows correspond to different methods of obtaining the quartets: NJ **(a)**, ML **(b)**, EMtree **(c)**, and the two bottom rows correspond to global methods: **(d)** global ML (estimating homogeneous continuous-time K81 model), **(e)** global NJ (with K81 distances). Columns correspond to the three different 12-taxon topologies simulated: *cc* (left), *cd* (middle), *dd* (right).

(a) QFIT with `NJ` quartets as input (600 bp).

(b) QFIT with `ML` quartets as input (600 bp).
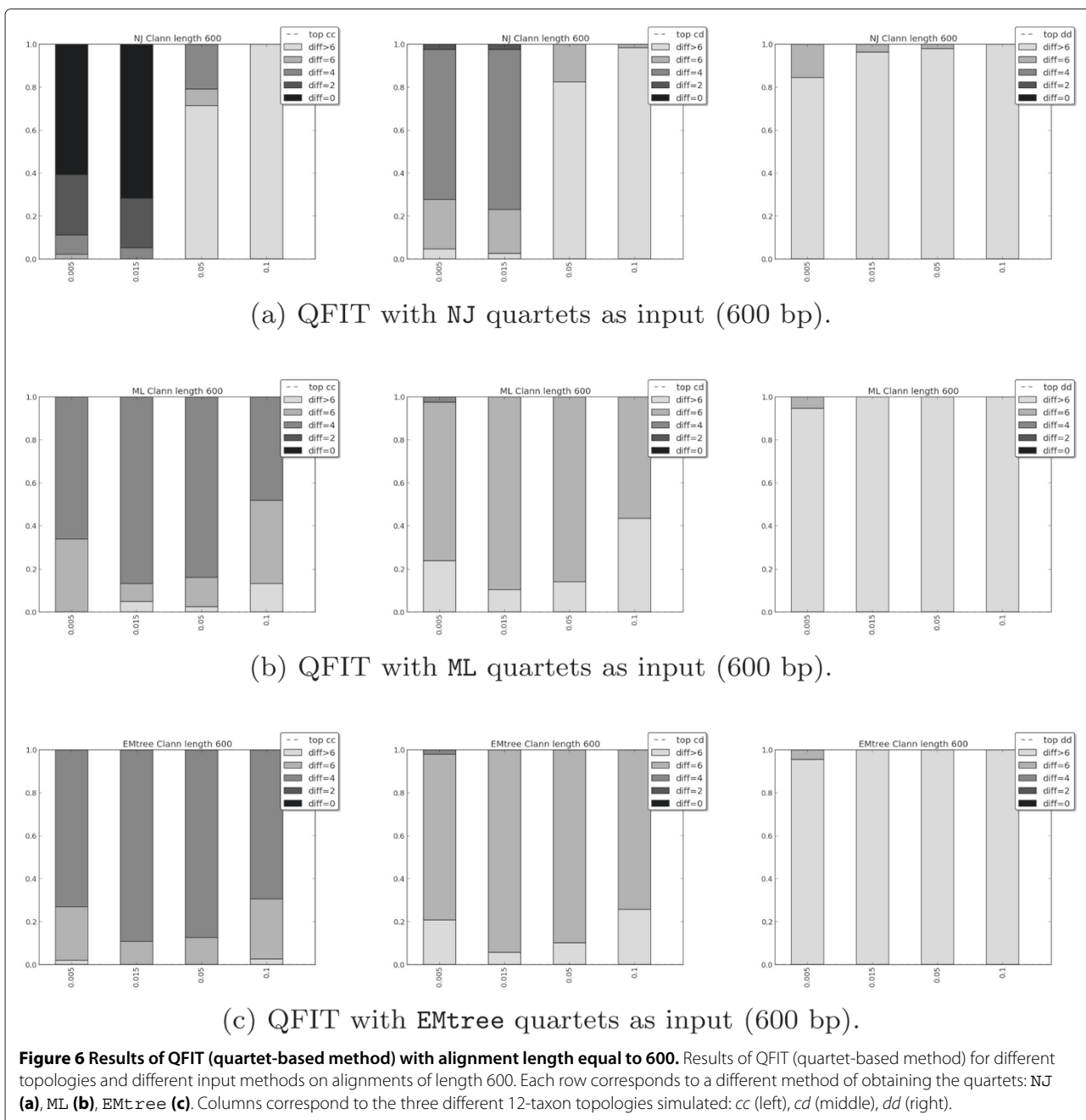
(c) QFIT with `EMtree` quartets as input (600 bp).

Figure 6 Results of QFIT (quartet-based method) with alignment length equal to 600. Results of QFIT (quartet-based method) for different topologies and different input methods on alignments of length 600. Each row corresponds to a different method of obtaining the quartets: `NJ` **(a)**, `ML` **(b)**, `EMtree` **(c)**. Columns correspond to the three different 12-taxon topologies simulated: *cc* (left), *cd* (middle), *dd* (right).

`EMtree` is a time-consuming algorithm, however, and the user has to decide whether it is worth performing such an analysis (we only recommend it for at most 6 taxa).

We have also assessed `EMtree`, `NJ`, and `ML` as input for the quartet-based methods QFIT and QuartetSuite. To do so, we have considered three different topologies on twelve taxa evolving under tome nonhomogeneous processes, and a wide set of branch lengths values. `EMtree` turns out to be the input method that performs best in most cases on this type of data. Regardless of the quartet-based method chosen and the tree topology, `EMtree`

gives the best results for trees with large divergence among taxa ($b = 0.1$).

Summing up, an EM approach on Markov models provides an accurate 4-taxon tree reconstruction method suited for data not known to satisfy homogeneity and very useful as input of quartet-based methods, specially for largely divergent trees. However, the method presented here is not valid for data violating the i.i.d. hypotheses, such as data with variation across sites (Gamma-rates, invariable sites, mixtures, and others), or dependency among sites. The method should be strongly modified

Table 1 Mean of Robinson-Foulds distance with QuartetSuite and global ML and NJ (variance shown in parentheses), 600 bp

Top.	Method	Mean Robinson-Foulds distane with QuartetSuite, 600 bp			
		b			
		0.005	**0.015**	**0.05**	**0.1**
cc	NJ	7.63 (10.82)	6.55 (18.32)	10.70 (13.04)	12.82 (9.57)
	ML	3.70 (4.91)	2.16 (1.46)	3.04 (1.42)	7.37 (19.97)
	EMtree	3.77 (4.14)	2.22 (2.33)	3.20 (1.10)	3.62 (3.27)
	global NJ	1.68 (3.42)	1.60 (0.64)	6.96 (10.76)	13.16 (1.45)
	global ML	**1.04** (3.08)	**0.0** (0.0)	**0.0** (0.0)	**3.10** (0.99)
cd	NJ	8.01 (14.01)	8.96 (11.61)	11.16 (12.54)	12.31 (10.48)
	ML	5.16 (3.56)	**3.34** (2.03)	**3.39** (1.59)	8.12 (19.28)
	EMtree	5.31 (3.96)	3.71 (2.82)	3.67 (1.61)	**4.38** (4.12)
	global NJ	5.80 (3.72)	5.56 (0.69)	8.44 (0.69)	13.12 (1.31)
	global ML	**5.04** (3.08)	4.0 (0.0)	4.0 (0.0)	6.60 (0.84)
dd	NJ	10.55 (9.53)	9.45 (9.39)	10.90 (13.95)	11.97 (10.91)
	ML	7.71 (3.67)	**6.42** (1.12)	6.08 (1.58)	9.58 (17.09)
	EMtree	7.58 (5.07)	6.87 (1.73)	**5.94** (1.44)	**6.60** (4.60)
	global NJ	9.24 (0.94)	9.60 (0.64)	10.4 (2.88)	13.08 (1.31)
	global ML	**7.48** (0.77)	8.0 (0.0)	8.0 (0.0)	10.60 (0.84)

in order to accommodate these generalizations, and this might be an interesting future project.

Methods

EM algorithm

We have implemented an expectation-maximization (EM) algorithm on four-taxon trees evolving under Markov processes. The core of the EM algorithm for phylogenetic trees is:

- *Input:* Data D (multiple alignment of n sequences), unrooted trivalent tree topology T with n leaves (and $\mathfrak{e} := 2n - 3$ edges), Markov model $\mathcal{M}$.

- *Initialization:* Root the tree at some internal node and provide tentative initial values for the root distribution π and the substitution matrices S_i ($i = 1, \ldots, \mathfrak{e}$), and a threshold $\varepsilon > 0$.
- *Recursion:*

  – *E-step:* Provide complete data cD (for all nodes in the tree) that maximizes the posterior probability $P_{T,\mathcal{M}}(S_i, \pi | cD)$ (unique maximum that can be computed efficiently by Felsenstein's algorithm [24]).
  – *M-step:* Compute the parameters $\hat{S}_i, \hat{\pi}$ that maximize the loglikelihood

Table 2 Mean of Robinson-Foulds distance with QFIT and global ML and NJ (variance shown in parentheses), 600 bp

Top.	Method	Mean Robinson-Foulds distance with QFIT, 600 bp			
		b			
		0.005	**0.015**	**0.05**	**0.1**
cc	NJ	**1.05** (2.22)	**0.67** (1.31)	8.12 (8.44)	12.97 (2.55)
	ML	4.68 (0.90)	4.36 (0.98)	4.37 (0.80)	5.31 (2.03)
	EMtree	4.58 (0.98)	4.21 (0.38)	**4.25** (0.44)	**4.66** (1.09)
cd	NJ	**4.60** (1.41)	**4.47** (1.11)	9.53 (5.68)	11.56 (2.46)
	ML	6.43 (0.88)	6.21 (0.37)	6.28 (0.48)	6.92 (1.18)
	EMtree	6.41 (0.98)	6.11 (0.21)	**6.20** (0.36)	**6.51** (0.76)
dd	NJ	**7.93** (1.11)	8.41 (1.16)	10.23 (2.51)	9.96 (2.89)
	ML	8.17 (0.74)	8.15 (0.27)	8.11 (0.20)	8.75 (1.14)
	EMtree	8.14 (0.63)	**8.05** (0.11)	**8.06** (0.11)	**8.44** (0.68)

$l(\{S_i\}_i, \pi) = P_{T,\mathcal{M},\{S_i\},\pi}(cD)$ for these complete data (this maximum is unique and can be computed in a closed form, see the Additional file 1, section 2).

– If $l(\{\hat{S}_i\}_i, \hat{\pi}) - l(\{S_i\}_i, \pi) > \varepsilon$, then set $\pi = \hat{\pi}$, $S_i = \hat{S}_i$ ($i = 1, \ldots, \mathfrak{e}$), and go back to *E-step*.

- *Output:* $\{\hat{S}_i\}_i$, $\hat{\pi}$ and $l(\{\hat{S}_i\}_i, \hat{\pi})$.

We set up the EM algorithm with $\varepsilon = 10^{-3}$ and forced it to stop after 100 iterations if it had not finished.

Given four aligned nucleotide sequences *s1, s2, s3, s4*, we run the EM algorithm for the three possible trivalent trees on four taxa: $T_1 = (s1, s2|s3, s4)$, $T_2 = (s1, s3|s2, s4)$, $T_2 = (s1, s4|s2, s3)$. Our tree reconstruction method returns the tree topology for which the likelihood output by EM is higher. This method is called `EMtree` throughout the paper.

Models

To test the method proposed in this paper, we mainly used the K81* model on four-leaved trees. This is, the evolutionary process is a Markov process with uniform distribution at the root and is specified by transition matrices of type

$$\begin{pmatrix} a_i & b_i & c_i & d_i \\ b_i & a_i & d_i & c_i \\ c_i & d_i & a_i & b_i \\ d_i & c_i & b_i & a_i \end{pmatrix}$$

on each branch e_i. In the matrix above, the rows and columns are labeled by nucleotides adenine, cytosine, guanine and thymine (in this order), so that entry (j, k) stands for the conditional probability that nucleotide j at the parent node of edge e_i is substituted by nucleotide k at the child node. When trees evolve under this model, the root becomes unidentifiable (different root locations may give rise to the same joint distribution at the leaves, [10]) and, as a consequence, one can only expect to reconstruct unrooted trees. This model does not assume a constant instantaneous mutation rate matrix over the tree and, therefore, trees evolving under this model are time nonhomogeneous. The more restrictive models K80* and JC* (Markov versions of Kimura 2-parameter and Jukes-Cantor models) are obtained by imposing $b_i = d_i$ and $b_i = c_i = d_i$, respectively ([10], § 4.2). All of them are time-reversible (and hence stationary), and are part of the so-called group-based models ([12,25]). The most general model among Markov processes is the *general Markov model* GMM, which considers transition matrices and root distribution without any further restriction and is neither stationary nor time-reversible [10].

The other two methods that are confronted to `EMtree` in this paper are Neighbor-Joining (`NJ`) and a usual continuous-time maximum-likelihood (`ML`). The ML estimates for homogeneous continuous-time K81 model [13] (with a instantaneous mutation rate matrix constant over the tree) have been obtained by the free software PAML [26] (`baseml` program), whereas for nonhomogeneous continuous-time K81 model we used the program bppml in the Bio++ package [22]. Neighbor-Joining was implemented considering the K81 distance [13]. When a global `ML` was used on 12 taxa, we used PAML and set up the stepwise addition option on this software.

Simulations

The simulated data in this paper has been produced using the program GenNon-h of [27] and Seq-Gen [28]. GenNon-h produces directly transition matrices of the required branch length (for any of the Markov models described above) and therefore does not assume time-homogeneity (not globally over the tree, nor locally at the edges). Seq-Gen was used to generate data evolving under continuous-time K81 model (both homogeneous and globally nonhomogeneous). In order to generate non-homogeneous data, we made the software generate various edges evolving at different instantaneous rates by recording ancestral sequences. Alternatively, the software *Hetero* [29] could be also used to generate (global) time nonhomogeneous continuous-time data.

Tree space on 4-leaf trees

In order to test the performance of the proposed method on four-taxon trees, we based our tests on the tree space proposed by Huelsenbeck [17], so that it is possible to compare our results to those obtained there with different phylogenetic reconstruction methods. In this tree space, two branch length parameters *a, b* on trees of four taxa are varied. Parameter *a* assigns the branch length to the internal branch and two opposite peripheral branches, and parameter *b* assigns the branch length to the two remaining branches as in Figure 2. Parameters *a* and *b* represent expected number of substitution per site and in this paper are varied from 0.01 to 0.75 in increments of 0.02.

For each pair (a, b), we simulated one hundred alignments of lengths 300 and 1000 basepairs (briefly bp) under the tree topology 12|34 (see Figure 2) and we inferred the topology using the three methods `EMtree`, `ML`, and `NJ`. The results for each method are shown in Figure 1.

Quartet-based methods

In order to assess the four-taxon tree reconstruction method `EMtree` proposed above, it is important to test its performance in quartet-based reconstruction methods. To do so, we considered two of these methods: Maximum Quartet Fit (QFIT) and the method proposed by Willson in [19]. QFIT choses the supertree that shares the maximum number of quartets with the source trees, whereas

the other method choses the supertree that optimizes a certain criterion based on the weights assigned at the input quartets. We use the implementation of QFIT distributed in Clann [21] and the implementation of Willson's method called QuartetSuite in [20]. While QuartetSuite produces a tree that is expected to minimize inconsistencies with quartets, QFIT uses heuristics to search through the tree space (and we restricted this search to 100000 trees). As our intention was not to evaluate quartet-based methods but testing our 4-taxon method in comparison to others, our criterion to choose these two quartet-based methods was the fact that they were freely available and that one of them allowed the use of weighted quartets.

In QuartetSuite weights are understood as -log of the frequency of a quartet, so that weights are nonnegative and zero denotes the tree with most support. Therefore we used the opposite of the loglikelihood output by `EMtree` and `baseml` as weights of the corresponding quartets. Weights for `NJ` were set up to be 0 for the topology output by `NJ` and 999 for the other two trees.

We followed [6] to test the performance of our method as input of quartet-based method. In that paper, the authors consider three tree topologies on 12 taxa, denoted as *cc*, *cd* and *dd* (see Figure 4), and fix the proportions among their branch lengths in order to compare different reconstruction methods. A parameter *b* denoting the internal branch lengths is varied between 0.005, 0.015, 0.05, and 0.1, which gives a maximum pairwise divergence along the tree of about 0.1, 0.3, 1.0, and 2.0 nucleotides per site, respectively. The lengths of the simulated alignments are 300 and 600 bp.

For each of these scenarios we generated 1000 alignments using GenNon-h and the Kimura 3-parameter model as explained above. For each alignment, the tree was estimated using QFIT and QuartetSuite with the four-taxon methods `EMtree`, `ML` and `NJ` as input. Then the Robinson-Foulds distance to the original tree (that is, the number of partitions that are present in one tree but not in the other) was computed using DendroPy symmetric_difference function [30]. The results are shown in Figures 5, 6, A1 and A2 (in the Additional file 1), and in Tables 1, 2, and S1, S2 (in the Additional file 1), and explained in the Results section.

Additional file

> **Additional file 1:** A pdf file containing figures A1, A2, and tables S1, S2 in section 1 and the explicit computation of the $M - step$ for K81* model in section 2.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
The first author implemented the methods, produced results, and wrote the paper. The second author had the idea, directed the work, implemented

methods and wrote the paper. Both authors read and approved the final manuscript.

Author details
[1]Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C - 08193 Bellaterra, Barcelona, Spain. [2]Departament Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Avda. Diagonal 647, Barcelona, Spain.

References
1. Kelchner SA, Thomas MA: **Model use in phylogenetics: nine key questions.** *Trends Ecol Evol* 2007, **22**(2):87–94.
2. Ripplinger J, Sullivan J: **Assessment of substitution model adequacy using frequentist and Bayesian methods.** *Mol Biol Evol* 2010, **27**(12):2790–2803.
3. Jermiin LS, Ho SY, Ababneh F, Robinson J, Larkum AW: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** *Syst Biol* 2004, **53**(4):638–643.
4. Galtier N, Gouy M: **Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis.** *Mol Biol Evol* 1998, **154**(4):871–879.
5. Yang Z, Yoder AD: **Estimation of the transition/transversion rate bias and species sampling.** *J Mol Evol* 1999, **48**:274–283.
6. Ranwez V, Gascuel O: **Quartet-based phylogenetic inference: improvements and limits.** *Mol Biol Evol* 2001, **18**(6):1103–1116.
7. Anderson FE, Swofford DL: **Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA.** *Mol Phylogenet Evol* 2004, **33**(2):440–451.
8. Semple C, Steel M: *Phylogenetics, Volume 24 of* Oxford Lecture Series in Mathematics and its Applications. Oxford: Oxford University Press; 2003.
9. Jayaswal V, Jermiin LS, Robinson J: **Estimation of phylogeny using a general Markov model.** *Evolutionary Bioinformatics Online* 2005, **1**:62.
10. Allman ES, Rhodes JA: **Phylogenetic invariants.** In *Reconstructing Evolution.* Edited by Gascuel O, Steel M. New York: Oxford University Press; 2007.
11. Barry D, Hartigan JA: **Statistical analysis of hominoid molecular evolution.** *Stat Sci* 1987, **2**(2):191–207.
12. Evans S, Speed T: **Invariants of some probability models used in phylogenetic inference.** *Ann Statist* 1993, **21**:355–377.
13. Kimura M: **Estimation of evolutionary distances between homologous nucleotide sequences.** *Proc Natl Acad Sci USA* 1981, **78**:454–458.
14. Dempster A, Laird N, Rubin D: **Maximum likelihood estimation from incomplete data via the EM algorithm.** *J Roy Stat Soc* 1977, **39**:1–38.
15. McLachlan G, Krishnan T: *The EM Algorithm and Extensions, Volume 382.* New York: Wiley-Interscience; 2007.
16. Kedzierska AM, Casanellas M: **EMpar: EM-based algorithm for parameter estimation of Markov models on trees.** [http://arxiv.org/abs/1207.1236]
17. Huelsenbeck J: **Performance of phylogenetic methods in simulation.** *Syst Biol* 1995, **44**:17–48.
18. Ho SY, Jermiin LS: **Tracing the decay of the historical signal in biological sequence data.** *Syst Biol* 2004, **53**(4):623–637.
19. Willson SJ: **Building phylogenetic trees from quartets by using local inconsistency measures.** *Mol Biol Evol* 1999, **16**:685–693.
20. Department of Computer Science, Iowa State University: **QuartetSuite by Raul Piaggio.** [http://genome.cs.iastate.edu/CBL/download/]
21. Creevey CJ, McInerney JO: **Clann: investigating phylogenetic information through supertree analyses.** *Bioinformatics* 2005, **21**:390–392.
22. Dutheil J, Boussau B: **Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs.** *BMC Evol Biol* 2008, **8**:255.

23. Strimmer K, Goldman N, von Haeseler A: **Bayesian probabilities and quartet puzzling.** *Mol Biol Evol* 1997, **14**(2):210.
24. Felsenstein J: *Inferring Phylogenies.* Sunderland: Sinauer Associates; 2004.
25. Szekely LA, Steel MA, Erdos P: **Fourier calculus on evolutionary trees.** *Adv Appl Math* 1993, **14**(2):200–216.
26. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *CABIOS* 1997, **15:**555–556. [http://abacus.gene.ucl.ac.uk/software/paml.html]
27. Kedzierska AM, Casanellas M: **GenNon-h: generating multiple sequence alignments on nonhomogeneous phylogenetic trees.** *BMC Bioinformatics* 2012, **13:**216.
28. Rambaut A, Grassly N: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997, **13:**235–238.
29. Jermiin LS, Ho SY, Ababneh F, Robinson J, Larkum AW: **Hetero: a program to simulate the evolution of DNA on a four-taxon tree.** *Appl Bioinformatics* 2003, **2**(3):159–163.
30. Sukumaran J, Holder MT: **DendroPy: a Python library for phylogenetic computing.** *Bioinformatics* 2010, **26:**1569–1571.