BIOINFORMATICS

Vol. 23 ECCB 2006, pages e136–e141
doi:10.1093/bioinformatics/btl304

# Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates

Matan Ninio[1,†], Eyal Privman[2,†], Tal Pupko[2,*] and Nir Friedman[2]

[1]The Selim and Rachel Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel and [2]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

**ABSTRACT**

Distance-based methods for phylogeny reconstruction are the fastest and easiest to use, and their popularity is accordingly high. They are also the only known methods that can cope with huge datasets of thousands of sequences. These methods rely on evolutionary distance estimation and are sensitive to errors in such estimations. In this study, a novel Bayesian method for estimation of evolutionary distances is developed. The proposed method enables the use of a sophisticated evolutionary model that better accounts for among-site rate variation (ASRV), thereby improving the accuracy of distance estimation. Rate variations are estimated within a Bayesian framework by extracting information from the entire dataset of sequences, unlike standard methods that can only use one pair of sequences at a time. We compare the accuracy of a cascade of distance estimation methods, starting from commonly used methods and moving towards the more sophisticated novel method. Simulation studies show significant improvements in the accuracy of distance estimation by the novel method over the commonly used ones. We demonstrate the effect of the improved accuracy on tree reconstruction using both real and simulated protein sequence alignments. An implementation of this method is available as part of the SEMPHY package.

**Contact:** talp@tau.ac.il

## 1 INTRODUCTION

The problem of phylogeny reconstruction is at the heart of evolutionary studies. Accurate knowledge of phylogenies is also instrumental in many tasks of protein sequence analysis, such as remote homology search (Altschul *et al*., 1997) and prediction of functional determinants in protein sequences (Nielsen, 1997; Pupko *et al*., 2002).

Several different approaches to phylogeny reconstruction have been developed over the past five decades, each with its own strengths and weaknesses. Maximum likelihood (ML) is a well-established methodology in phylogeny reconstruction. ML methods use a stochastic model of sequence evolution that describes the probabilities of substitutions. The ML estimate (MLE) for the phylogeny is the tree that maximizes the conditional probability of the sequence data, given this tree and the model. This probability is called the likelihood of the data. ML methods have been argued to be superior in terms of accuracy and statistical justification (Fukami-Kobayashi and Tateno, 1991; Hasegawa, 1993; Kuhner and Felsenstein, 1994; Tateno *et al*., 1994; Huelsenbeck, 1995).

However, ML methods become computationally infeasible when dealing with large datasets because the tree search space, i.e. the number of possible trees, grows exponentially with the number of sequences (Felsenstein, 2004). The currently available applications of ML methods to phylogeny cannot effectively cope with more than a few hundreds of sequences. This problem becomes increasingly aggravating with the rapid accumulation of molecular sequence data. In many molecular studies it is now possible to compile a dataset of hundreds and even thousands of homologous sequences. Concomitantly, the field of molecular evolution has produced increasingly sophisticated methods for phylogenetic analysis, which are more computationally intensive. These combined advances challenge contemporary studies of molecular evolution.

Contrary to ML methods, the efficiency of distance-based methods, which are discussed in this article, is polynomial in terms of the number of sequences. This advantage in computation time makes them essential for dealing with large datasets. The importance of distance methods is not only as a faster, less accurate alternative to ML methods, but also in providing a good starting point of a heuristic search for the ML tree. (Friedman *et al*., 2002; Guindon and Gascuel, 2003). Clearly, if the distance method could be improved then the ML search could be faster, and give more accurate results.

Distance-based methods are made up of two steps:

(1) Pairwise distance estimation between all possible pairs of sequences in the dataset.

(2) Tree reconstruction based on the distances only. This stage does not use the original sequences.

These are the two modular stages—any method for distance estimation can be used with any distance-based method for tree reconstruction. While several distance-based tree reconstruction methods have been developed, the initial step of distance estimation received scant attention. Indeed, the simplistic Jukes–Cantor (JC) method (Jukes and Cantor, 1969) is still a common practice for distance estimation, in spite of its oversimplifying assumption that all types of substitutions have equal probabilities. Great efforts have been invested in improved modeling of sequence evolution for use with ML methods. These improvements should also be applied to distance estimation. However, as this work will demonstrate, all the

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

 

previously published distance methods are still inadequate in terms of both error and bias.

We present a novel approach to distance estimation with increased accuracy, thereby improving phylogeny reconstruction. Our method is an adaptation of advanced ML models to a distance-based approach, making them computationally feasible for many thousands of sequences. Thus, the analysis of large datasets may now benefit from the improved accuracy of these fine-tuned models. The key idea underlying our novel methodology is to extract information from the entire dataset, and to use it in each pairwise distance estimation. We show that the novel method significantly improves the accuracy of phylogenetic tree reconstruction. This method was implemented as part of the SEMPHY package (Friedman *et al*., 2002). Source code and executable are available for download at http://compbio.cs.huji.ac.il/semphy/

## 2 METHODS FOR ML DISTANCE ESTIMATION

The evolutionary distance, $d$, between a pair of sequences is defined as the average number of substitutions per sequence site. This measure is related to the time that passed and the rate of substitutions. ML methodology may be used for distance estimation in a similar fashion to finding the ML tree: The MLE for the distance between two sequences is the distance that maximizes the likelihood of the observed sequences (Zharkikh, 1994). ML methods use continuous time Markovian models (Karlin and Taylor, 1975) that define for any pair of aligned characters $a$ and $b$ the probability $p_{a\to b}(d)$ of the substitution from $a$ to $b$ in an evolutionary distance $d$. In addition, they define the initial character probabilities $\pi(a)$. These models range from the oversimplified JC model (Jukes and Cantor, 1969) to complex models that strive to capture the nature of evolution of protein-coding genes as accurately as possible. This section describes the cascade of ML methods of increasing complexity, culminating in the novel methods we propose.

### 2.1 Assuming homogeneous substitution rates

All distance estimation methods that are in wide-spread use for distance-based tree reconstruction assume no variation of the substitution rates between different sites. The basic JC distance method assumes uniform frequencies of all characters and equal probabilities for all substitutions. These unrealistic assumptions can be avoided by employing various substitution matrices, which are used to calculate the $p_{a\to b}(d)$ probabilities. Such matrices have been initially designed for nucleotides sequences (Kimura, 1980). For amino acids, the larger alphabet size (20 instead of 4) requires a significantly larger number of parameters in the model. Therefore, empirical replacement matrices were calculated using large protein datasets. In this work we concentrate on amino acid sequences, for which the computational challenge is greater, although our novel methods can be equally applied to DNA sequences. Specifically, we use the JTT matrix (Jones *et al*., 1992).

Contrary to the JC distance method, there is no closed formula for the ML distance when using these matrices. Therefore, the likelihood of the data is maximized using numerical methods. Under the simplifying assumption that sites evolve independently and for models that satisfy reversibility $[\pi(a)p_{a\to b}(d) = \pi(b)p_{b\to a}(d)]$ (Felsenstein, 2004), the likelihood of a pair of sequences can be written as

$$L(d) = P(A,B \mid d) = \prod_{i=1}^{S} \pi(a_i) P_{a_i \to b_i}(d) \qquad (1)$$

where $a_i$ and $b_i$ are the characters in the $i$-th positions in sequences $A$ and $B$ respectively, out of a total of $S$ positions in the sequence alignment. The most significant oversight of this model is the assumption of equal replacement rates at all amino acid sites. In this article, we shall refer to the method that uses this model as the *homogeneous rates* method. However, evolutionary rates vary considerably between different amino acid sites, owing to non-uniform selection forces (Yang, 1996).

### 2.2 Among site rate variation

Models that explicitly take into account among-site rate variation (ASRV) were shown to be statistically superior to the homogeneous models (Yang, 1994). ASRV is modeled by assuming that each site $i$ in the sequence has a different rate, $r_i$, relative to the average rate over all sites. Thus, a site of rate 2 evolves twice as fast as the average. This is equivalent to multiplying the distance by the rate in the likelihood calculation for each site:

$$L(d) = P(A,B \mid r,d) = \prod_{i=1}^{S} \pi(a_i) P_{a_i \to b_i}(d \cdot r_i) \qquad (2)$$

This equation assumes that rates are known. Since this is not the case, a prior distribution of rates $R(r)$ is assumed. The likelihood is then computed by averaging over all possible rates:

$$L(d) = P(A,B \mid R,d) = \prod_{i=1}^{S} \int_{r=0}^{\infty} R(r)\pi(a_i) P_{a_i \to b_i}(d \cdot r)\mathrm{d}r. \qquad (3)$$

The most common choice for $R(r)$ is the gamma distribution with the mean set to 1 (Yang, 1996). The gamma density function has one free parameter, $\alpha$, that allows for different distribution shapes. The distance and the $\alpha$ parameter can be estimated simultaneously for each pair of sequences, using ML. We shall refer to this method as the *pairwise* $\alpha$ method. most practical situations a discrete approximation of the gamma distribution is used. Here we use 32 discrete bins.

### 2.3 Iterative inference of model parameters

The *pairwise* $\alpha$ method estimates the $\alpha$ parameter for each pair of sequences independently. However, the variability of rates in a protein is generally common to all sequences across a given multiple sequence alignment (MSA). Thus, there is no reason to estimate the rate parameters for each pair of sequences. Moreover, such estimation of many parameters from scant data is likely to result in high errors (Fig. 1a). It would be preferable to use all sequences in order to estimate the rate parameters globally. However, such estimation requires knowledge of the phylogenetic tree, which we have not yet reconstructed. This kind of circular situation calls for an iterative process of optimization. Sullivan *et al*. (2005) studied iterative parameter optimization in the context of ML tree search. Here we suggest a similar approach for distance-based tree reconstruction. Global ASRV information is extracted from the entire MSA, using the tree reconstructed in the previous iteration. This 'global information' is then used to re-estimate the pairwise distances more accurately (Fig. 1b).

There are various alternatives in the estimation of ASRV parameters:

*Iterative* $\alpha$: Initial pairwise distances are estimated using the *homogeneous rates* method, and a tree is reconstructed. This tree is used to infer $\alpha$, and $\alpha$ is then used to improve the estimation of the pairwise distances. These iterations are repeated until the likelihood converges.

*Iterative rates*: This method uses the evolutionary rate at each position as the 'global information'. The MLEs of these rates are iteratively estimated, and then used to recalculate the distances by maximizing Equation (2). This method captures more information about the ASRV than the *iterative* $\alpha$ method.

*Iterative* posterior: Here we propose a third alternative that calls for inferring a (posterior) rates distribution for each site rather than relying on a single estimate of the ML rate. This distribution is then used in Equation (3) instead of the prior distribution $R(r)$. In the discrete approximation that is used here, the posterior probability of each rate category is calculated for each site, in each iteration, as described in Equation (2) in Mayrose *et al*. (2004). We show that this novel approach outperforms all other methods.

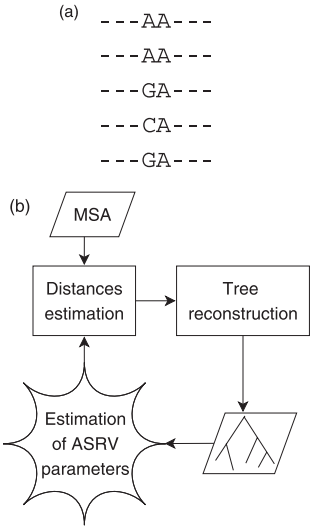


**Fig. 1.** Utilizing the entire MSA to estimate the variation of rates between sites. (**a**) When looking only at the first two sequences of this simple example, both sites of the alignment are identical and there is no reason to think that they evolve with different rates. However, when all five sequences are used, we can deduce that the rate of the first site is larger than that of the second one. (**b**) The proposed iterative approach that utilizes ASRV information from all sequences to improve distance estimation.

## 3 EVALUATION OF THE DISTANCE ESTIMATION METHODS

The performance of the different methods was evaluated in three comparative studies. The results presented here are for the five methods summarized in Table 1. In addition, all studies included the JC distance method. However, these results are not shown because of its poor performance compared with all other methods. Its error and bias patterns follow those of the *homogeneous rates* method.

### 3.1 Reconstructing trees from protein sequence alignments

The ultimate goal of improving distance estimation is to increase the accuracy of the reconstructed tree topology. Therefore, the accuracy of reconstruction using the novel methods was evaluated both for real and simulated protein sequences. We used the neighbor joining (NJ) method for tree reconstruction (Saitou and Nei, 1987), which is the most popular distance-based method. Nevertheless, our novel distance estimation methods can be equally used with other distance-based methods, which have been shown to be superior to NJ in certain aspects (Huson *et al*., 1999).

We used a dataset of 84 protein MSAs that was composed by Aloy *et al*. (2001). For each MSA, NJ trees were reconstructed using the five different distance methods, and compared in terms of their log-likelihood scores. In order to produce comparable scores, all likelihoods were computed using the gamma ASRV model. In addition, such comparison might be affected by biases in branch length estimation under the different models. Therefore, branch lengths and $\alpha$ optimization were performed on the fixed tree topologies that were constructed by NJ. Each log-likelihood score was divided by the length of the MSA to produce the average log-likelihood score per position. Table 2 lists the differences between the score of each method and the score of the *homogeneous rates* method, which is used as a reference. The second line indicates the number of MSAs for which there was a difference in the tree topology that resulted in improved likelihood, compared with the *homogeneous rates* method.

**Table 1.** The distance estimation methods used in all comparative Studies

| Name | Evolutionary model |
|---|---|
| *Homogeneous rates* | No ASRV |
| *Pairwise* $\alpha$ | Independent estimation of $\alpha$ for each sequence pair |
| *Iterative* $\alpha$ | Global estimation of $\alpha$ |
| *Iterative rates* | Global estimation of the ML rate at each site |
| *Iterative posterior* | Global estimation of the posterior distribution of the rate at each site |

**Table 2.** Tree reconstruction by different methods

| Average | *Pairwise* $\alpha$ | It. rates | It. $\alpha$ | It. posterior |
|---|---|---|---|---|
| Per position[†] $\Delta$LL | −0.0655 | +0.0151 | +0.0077 | +0.0177 |
| Improved topology[‡] | 7% | 31% | 32% | 44% |

[†]The average difference in the log-likelihood per position scores compared to the *homogeneous rates* method.
[‡]The proportion of trees for which there was a difference in the topology and an improved likelihood compared to *homogeneous rates*.

Compared with this reference, the *pairwise* $\alpha$ method produces trees of lower likelihood for most cases. On the other hand, all three *iterative* methods improve the likelihood scores on average. The *iterative* posterior method achieved the best results, with an average improvement of 0.0177 log-likelihood points per position and an improved topology for 44% of the MSAs. We used simulation studies to further investigate this pattern.

### 3.2 Reconstructing trees from simulated multiple sequence alignments

Accuracy of tree reconstruction from real protein sequences can only be compared in terms of the likelihood of the trees, since the true phylogeny is not known. For this reason we applied the different methods to protein MSAs that were simulated according to a known tree, and we evaluated their accuracy by comparing the reconstructed tree with the original tree. We used 10 trees that were reconstructed by the *homogeneous rates* method in the previous section as the basis for the simulated MSAs. Thus these simulations represent several tree topologies of real protein phylogenies. We chose MSAs with a number of sequences around 50.

The gamma-ASRV model was used to simulate sequence evolution according to those tree topologies. The simulations were repeated for ten values of $\alpha$: 0.1 (highly variable rates), 0.2, 0.5, 0.7, 1.0, 1.3, 1.6, 2.0, 2.5 (relatively *homogeneous rates*). For each $\alpha$, a vector of 1000 rates was sampled from the gamma distribution. Each of the ten trees was used with each of the ten rate vectors to simulate an MSA of 1000 columns. This procedure was repeated ten times, resulting in ten MSAs for each tree and for each $\alpha$ value, a total of 1000 simulated MSAs. Each method was used to reconstruct a tree from each MSA and the resulting trees were compared to the original tree that was used to simulate this MSA.

The performance of the five methods was evaluated in terms of log-likelihood scores (as above) and in terms of the topological distance between the inferred and the original tree, namely the percent of splits that both trees agree on. Figures 2a and 2b plot these two accuracy measures as a function of $\alpha$ (presented in log-scale). Both measures agreed on the ranking of the five methods: *iterative posterior* > *iterative rates* > *iterative* $\alpha$ > *homogeneous rates* > *pairwise* $\alpha$. Paired *t*-test comparisons indicate these differences are highly significant (*p*-values lower than $10^{-5}$).

The results for the simulated MSAs agree with the pattern that was observed for the real protein sequences. The differences in the log-likelihood per position are also comparable. Therefore, we conclude that the above procedure successfully reproduces the general pattern of sequence evolution. An interesting observation is that *pairwise* $\alpha$ preforms especially bad for simulations with extreme values of $\alpha$, and is therefore worse than *homogeneous rates*. This is the result of large errors in the $\alpha$ estimates, which are based on two sequences only (data not shown).

Compared to the commonly used *homogeneous rates* method, the *iterative posterior* method improves the log-likelihood score by 0.01-0.03 points per position, depending on $\alpha$. In terms of the topological accuracy of the tree, the percentage of correctly reconstructed splits is improved by 2%–9%, depending on $\alpha$. A larger improvement is evident for $\alpha$ values less than 1. Therefore, this novel method will be especially significant for proteins with large rate heterogeneity. The improvement in correct split reconstruction is usually very valuable, as we observed that many of the longer branches are easily reconstructed with any distance estimation method, and a relatively small number of short branches is commonly the more challenging part of the phylogeny. This pattern is plotted in Figure 2c. The largest impact is on branch lengths around 0.01, where the proportion of correctly reconstructed splits is improved by 20%.

## 3.3 Evaluation of the accuracy of distance estimation on pairs of simulated sequences

The evaluation of tree reconstruction above clearly shows the superiority of the *iterative* methods. However, it is interesting to understand how the improvements in the accuracy depend on different factors, such as the pairwise distances and the $\alpha$ parameter. For example, improvements in the accuracy for relatively distant pairs of sequences might be more significant than for close pairs. In addition, the different methods may vary in the extent of their bias in distance estimation. Therefore, we used simulations of pairs of sequences to study the effects of these factors. We investigated the error and the bias by comparing the estimated distance with the original distance that was used in the simulation.

The same protocol that was used to simulate MSAs was adapted to simulate pairs of sequences 1000 amino acids long. 1000 pairs were simulated for each combination of the 10 different $\alpha$ values and 10 different evolutionary distances between 0.01 and 1.5. In total, 100,000 pairs of sequences were simulated. For the *iterative* methods we used the previously simulated MSAs in order to estimate the required 'global information', i.e., the global $\alpha$ parameter, and for each site—the ML rate and the posterior distribution of the rate. For each pair we used an MSA that was simulated with the same rate vectors, so that the new sequence pair can be treated as though it belongs to the same dataset.

The accuracy of the five distance methods was evaluated on these simulations. In addition, a sixth method (labeled *true rates*) was added as a frame of reference. This method is similar to the *iterative rates* method, however it was given the true rates that were used to simulate the sequences instead of the MLEs of the rates. This information is obviously not available for real proteins. It is used here in order to demonstrate the limit of the accuracy of this class of ML methods, when given the most accurate 'global information' about the rates.

The results were analyzed in terms of the error and the bias in distance estimation. The relative mean square error (RMSE) and the relative mean error (RME) were used to measure the error and the bias respectively:

$$Avg\left(\left(\frac{\hat{d}-d_{true}}{d_{true}}\right)^2\right) \quad Avg\left(\frac{\hat{d}-d_{true}}{d_{true}}\right) \tag{4}$$

*3.3.1 Accuracy as a function of the evolutionary distance* In Figure 3 the RMSE and RME of each method are plotted as a function

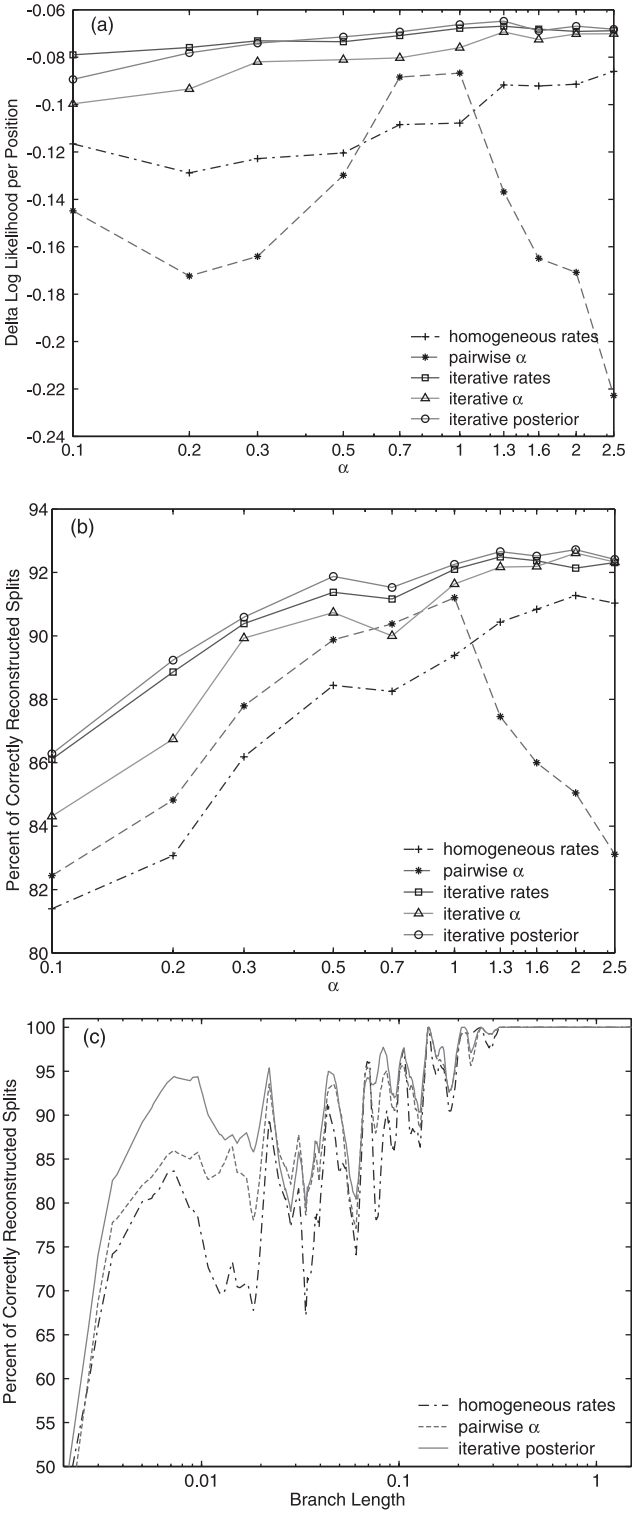


**Fig. 2.** Accuracy of tree reconstruction using the different distance estimation methods, plotted vs. the $\alpha$ value that was used in the simulations. (**a**) The difference in the log-likelihood per position of the reconstructed tree, compared to the true tree. (**b**) The percentage of split agreement with the true tree. (**c**) Percent of correctly reconstructed splits vs. the corresponding branch length. The curves were created using the LOWESS function (locally weighted scatter plot smooth).

of the true distance by which the sequence pairs were simulated. The results are shown for simulations with an $\alpha$ value of 0.7.

The improved accuracy of the novel *iterative* methods is evident from Figure 3a, especially for large distances. It seems that only at large distances, where many sites undergo multiple replacements, there is a significant advantage to the more refined models. For small distances most methods produce very similar errors. For distances larger than 0.2 all the ASRV methods are significantly more accurate than the *homogeneous rates* method. The major contributing factor to the inaccuracy of *homogeneous rates* is probably its considerable bias for underestimation (Fig. 3b), which increases dramatically with the distance.

Among the ASRV methods, the *iterative* methods that use 'global information' are significantly more accurate than the *pairwise gamma* method that do not. We attribute this result to the insufficiency of the information in two sequences for accurate estimation of ASRV parameters. Interestingly, there is a noticeable bias for overestimation (over 10 percent) in the *pairwise gamma* method, for both small and very large distances. The *iterative* methods, on the other hand, do not display a significant bias. The *iterative posterior* method seems to be especially unbiased.

As expected, the accuracy of all the methods never exceeds that of the *true rates* method, as the true rates are the ultimate 'global information'. Surprisingly, even for very large distances, the three *iterative* methods produce RMSE values that are no more than 1.5 times larger than the those of the *true rates* reference. In general, the *iterative posterior* method is more accurate than the other two methods. Its advantage is especially noticeable for large distances, where its errors are almost equal to the gold standard set by *true rates*.

It is worthwhile to note the effect of the improved accuracy of pairwise distances on the successful reconstruction of tree topology. The most significant improvement in the pairwise accuracy was for distant pairs (distances larger than 0.2), while the improved reconstruction was mainly in the shortest branches of the trees (of length around 0.01, as shown in Fig. 2c). Evidently, the accurate estimation of large pairwise distances is essential for resolving difficult splits that correspond to short branches. This effect is reasonable, because distant pairs of sequences are often connected by a path in the tree that includes very short branches. Therefore, the large pairwise distances are used by NJ to resolve those internal branches.

*3.3.2 Accuracy as a function of* $\alpha$ When ASRV models are applied to protein sequences the estimated $\alpha$ values typically range between 0.5 and 3.0. In order to test the effect of the degree of rate variation on the accuracy of the distance estimation methods we plotted the error and the bias against $\alpha$. Figure 4 presents the results for a distance of 1.0, which is a large but not uncommon distance. At most of the biologically relevant $\alpha$ values the three *iterative* methods are clearly more accurate than the simpler methods. However, at $\alpha$ values of 0.5 and smaller the *iterative posterior* and the *iterative rates* methods become less accurate, while the *iterative* $\alpha$ method remains nearly as accurate as the *true rates* reference.

This increased error is correlated with a bias for underestimation (Fig. 4b). We investigated the cause of this bias, finding that it was preceded by underestimation in the branch lengths of the trees that were reconstructed from the simulated MSAs. The bias of the ML estimation of the branch lengths at small $\alpha$ values was never reported before. This is an interesting and important result in itself, which merits further investigation, as it surely affects any other evolutionary analyses that make use of the branch lengths of trees. In our analysis, the shortening of the branch lengths resulted in overestimation of the rate at each site, which caused underestimation of distances by *iterative rates* and *iterative posterior*. Nevertheless, the novel methods we present here produce high accuracy at all situations except for the very extreme end of the rate variability in real biological protein sequences.

## 4 SUMMARY

The *JC distance* method is still in wide-spread use in distance-based phylogeny reconstruction. Even the more advanced methods that use amino-acid replacement matrices neglect to take ASRV into consideration. Thus, such methods suffer from high errors and bias, as we show in our simulation studies (sections 3.2 and 3.3). These simulations also show that an attempt to estimate ASRV parameters for each pair of sequences independently will inevitably suffer from large errors. Therefore, we use iterative tree reconstruction to extract more refined 'global ASRV information' from the entire dataset, using the tree that was estimated in the previous iteration. This 'global information' is utilized by a novel Bayesian distance estimation method that integrates the posterior distribution of the rate at each site into the estimation of the distance.

We demonstrate the improved accuracy of our novel method through a comparative study of distance estimation methods and their use in NJ. The novel *iterative* method produces trees of significantly improved likelihood for both real and simulated protein MSAs. The simulations also show that the novel method correctly reconstructs a larger percentage of the branches of the true tree, therefore, giving a better estimate of the tree topology. Using

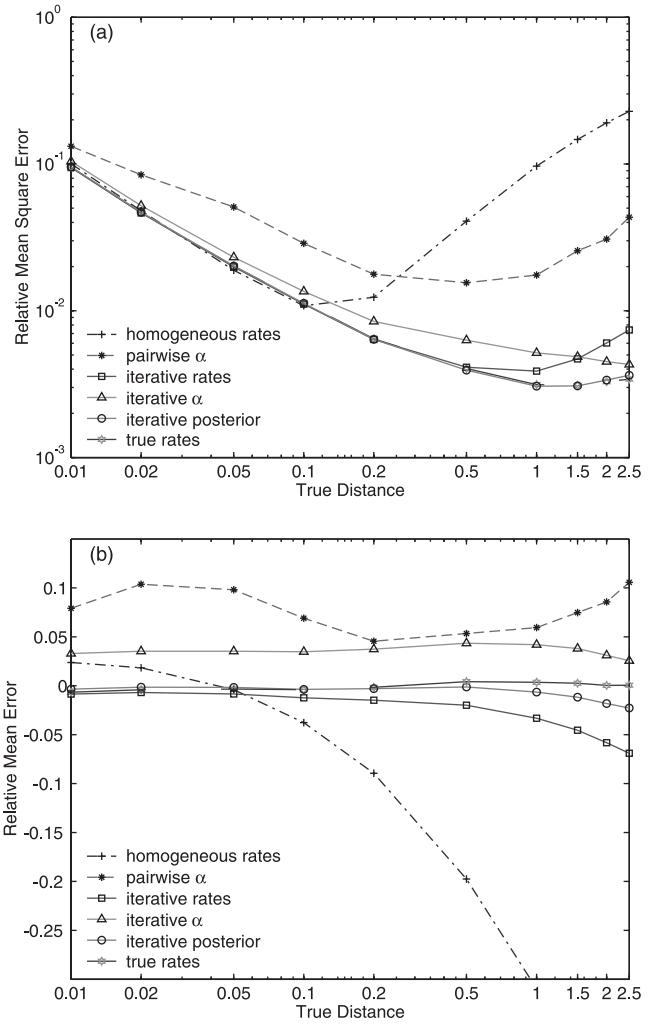


**Fig. 3.** Error and bias of the different distance estimation methods as a function of the true distance. Sequences were simulated with $\alpha = 0.7$. Each data point is an average based on 1000 independent sequence pairs. (**a**) RMSE as a measure of the error. (**b**) RME as a measure of the bias.

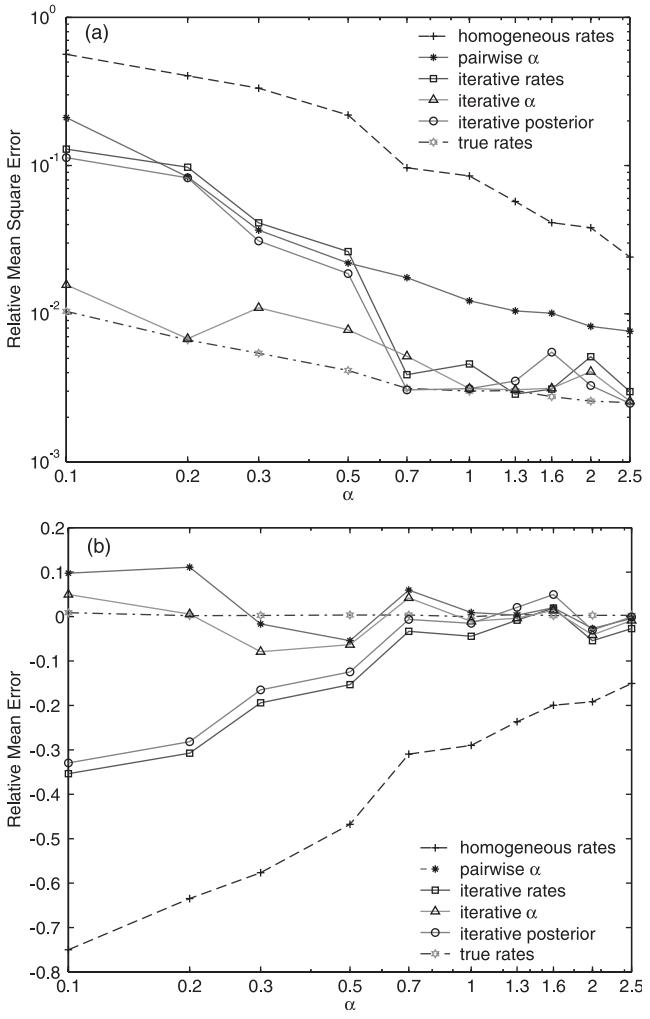


**Fig. 4.** Error and bias of the different distance estimation methods as a function of $\alpha$. Sequences were simulated with a pairwise distance of 1.0. Each data point is an average based on 1000 independent sequence pairs. (**a**) RMSE as a measure of the error. (**b**) RME as a measure of the bias.

simulations of sequence pairs we show that the 'global information' that is available to the *iterative* method reduces errors and bias in distance estimation. While all previously suggested distance-based methods consider each pair of sequences separately, the *iterative* method makes use of all available sequences, allowing a more accurate estimation of the parameters of the gamma-ASRV model. Our simulations demonstrate that these advantages are considerable in almost all cases, and are increasingly significant for large evolutionary distances and for proteins of high rate variability.

## ACKNOWLEDGEMENTS

This work was supported in part by an Israeli Science Foundation grant number 1208/04. TP was supported by a grant in Complexity Science from the Yeshaia Horvitz Association, and by a grant from the Israel Ministry of Science.

## REFERENCES

Aloy,P. *et al.* (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**(2), 395–408.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein,J. *et al.* (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Friedman,N. (2002) A structural EM algorithm for phylogenetic inference. *J Comput Biol*, **9**(2), 331–353.

Fukami-Kobayashi,K. and Tateno,Y. (1991) Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J Mol Evol*, **32**(1), 79–91.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**(5), 696–704.

Hasegawa,M. and Fujiwara,M. (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol Phylogenet Evol*, **2**(1), 1–5.

Huelsenbeck,J.P. (1995) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol Biol Evol*, **12**(5), 843–849.

Huson,D.H. *et al.* (1999) Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology*, **6**(3–4), 369–386.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, **8**, 275–282.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In H.N. Munro, editor, *Mammalian protein metabolism,* pp. 21–132. Academic Press, New York.

Karlin,S. and Taylor,H.M. (1975) *A first course in stochastic processes*. New York: Academic Press.

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.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, **11**(3), 459–468.

Mayrose,I. *et al.* (2004) Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Mol Biol Evol*, **21**(9), 1781–1791.

Nielsen,R. (1997) Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst Biol*, **46**(2), 346–353.

Pupko,T. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**(Suppl 1), pp. 71–77.

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

Sullivan,J. *et al.* (2005) Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. *Mol Biol Evol*, **22**(6), 1386–1392.

Tateno,Y. *et al.* (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol Biol Evol*, **11**(2), 261–277.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol*, **39**(3), 306–314.

Yang,Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, **11**, 367–372.

Zharkikh,A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, **39**(3), 315–329.