Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction

Volker Hollich, Lena Milchert,* Lars Arvestad,† and Erik L. L. Sonnhammer**

*Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden; and †Stockholm Bioinformatics Center, Albanova, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden

Distance-based methods are popular for reconstructing evolutionary trees of protein sequences, mainly because of their speed and generality. A number of variants of the classical neighbor-joining (NJ) algorithm have been proposed, as well as a number of methods to estimate protein distances. We here present a large-scale assessment of performance in reconstructing the correct tree topology for the most popular algorithms. The programs BIONJ, FastME, Weighbor, and standard NJ were run using 12 distance estimators, producing 48 tree-building/distance estimation method combinations. These were evaluated on a test set based on real trees taken from 100 Pfam families. Each tree was used to generate multiple sequence alignments with the ROSE program using three evolutionary models. The accuracy of each method was analyzed as a function of both sequence divergence and location in the tree. We found that BIONJ produced the overall best results, although the average accuracy differed little between the tree-building methods (normally less than 1%). A noticeable trend was that FastME performed poorer than the rest on long branches. Weighbor was several orders of magnitude slower than the other programs. Larger differences were observed when using different distance estimators. Protein-adapted Jukes-Cantor and Kimura distance correction produced clearly poorer results than the other methods, even worse than uncorrected distances. We also assessed the recently developed Scoredist measure, which performed equally well as more complex methods.

Introduction

The construction of phylogenetic trees has many applications in current biological research. It is a popular means to address evolutionary questions in, e.g., taxonomy, protein function inference, and epidemiology. Reconstruction of phylogenetic trees from the wealth of gene sequences has attracted many researchers over the years and has given rise to a large number of methods such as neighbor-joining (NJ), maximum likelihood (ML), minimum evolution (ME), and parsimony (see e.g., Nei 1996; Zhang and Nei 1997; Whelan, Lio, and Goldman 2001).

ML is often considered the best approach. Here, each tree topology is assigned a likelihood, summing over all possible ancestral sequences (Felsenstein 1981; Kishino, Miyata, and Hasegawa 1990). This is repeatedly carried out for all tree topologies, and the tree with the highest likelihood is finally chosen. The major drawback of ML is its poor scalability. Already with a small number of sequences, it becomes unfeasible to examine every possible tree topology. Therefore, the much faster distance-based methods have gained popularity and are today most widely used.

The simplest distance-based method, unweighted pair-group method using arithmetic averages (UPGMA), dates back to 1958 in its earliest version (Sokal and Michener 1958; Sneath and Sokal 1973). This method simply joins the two nodes with the shortest distance at each stage of the clustering. The newly formed cluster is given a distance to the remaining nodes calculated as arithmetic averages. A drawback with this approach is that all distances from the leaves to the root become the same. If the assumption of a molecular clock holds, this approach delivers acceptable topological results. However, UPGMA is known to yield poor results when substitution rates vary and is not trusted in general for phylogenetic tree construction.

The widely used NJ algorithm (Saitou and Nei 1987) performs clustering in a similar fashion as UPGMA, but it explicitly takes variable evolutionary rates into account. Atteson (1997) showed that if the error of the distance estimates is at most half the length of the shortest branch in the underlying phylogeny, then NJ always returns the correct tree. It has been shown that the topology of the NJ tree is close to that of the ME tree (Saitou and Nei 1987; Saitou and Imanishi 1989). Fast versions of NJ have been published (Howe, Bateman, and Durbin 2002; Mailund 2003) in which heuristics are used to avoid unnecessary recomputations.

BIONJ is a modification of NJ (Gascuel 1997) that uses a simple model of the sampling noise (variance) of evolutionary distances. During each step of the clustering process, nodes are selected for joining so that the variance of the new distance matrix is minimized. It thus takes into account the fact that high distances present a higher variance than short ones.

Another modification of NJ is the Weighbor algorithm or "weighted NJ" (Bruno, Socci, and Halpern 2000). Here, the selection of nodes for joining is based on "additivity" and "positivity" properties, which are estimated using ML. It is claimed to achieve tree accuracies comparable to exhaustive ML, yet at a much lower computational cost.

Besides NJ and its various modifications, other attempts to fulfill the ME criterion have been taken by Desper and Gascuel (2002). Their greedy minimum evolution algorithm is used to calculate a tree, which is further improved by nearest neighbor interchange. The authors presented unweighed and weighted versions of their approach, both implemented in their program FastME.

The starting point for all distance-based tree construction algorithms is the pairwise sequence distances. Several methods to estimate these are available. Some methods were originally formulated for DNA sequences but have

Key words: protein distance estimation, phylogenetic tree reconstruction, neighbor-joining.

E-mail: volker.hollich@cgb.ki.se.

Mol. Biol. Evol. 22(11):2257–2264. 2005
doi:10.1093/molbev/msi224
Advance Access publication July 27, 2005

been adapted to protein sequences, which are discussed in this study.

In general, the various distance estimation methods can be divided into two classes. In the first class, distance between two sequences is measured by simply calculating the percentage nonidentical sites. The distance is usually found by applying some mathematical correction to compensate for multiple substitutions (that cannot be observed directly) to this value.

In this paper, we evaluated the methods Kimura (1983) and uncorrected distance (percent differences) in this category. The Jukes-Cantor (1969) estimator has originally been formulated for DNA sequences. We used the Jukes-Cantor–inspired model by Takezaki, Rzhetsky, and Nei (1995) ($d = -(19/20)\ln(1-(20/19)\hat{p})$, where $\hat{p}$ is the estimate of the expected proportion [p] of sites that is different between the two sequences) and refer to is as Jukes-Cantor protein (JCP). We also tested a new method called Scoredist that is based on a logarithmic correction function applied to the alignment score according to the BLOSUM62 (S. Henikoff and J. G. Henikoff 1992) matrix (Sonnhammer and Hollich 2005). The principle is similar to the Kimura correction but is based on alignment score instead of percent identity. Scoredist needs to be calibrated with an evolutionary model; throughout this paper, we used the default Scoredist version which is calibrated using the Dayhoff model and default indels in ROSE (calibration factor = 1.3370).

Distance measures belonging to the second class use a series of amino acid substitution matrices. Here, the distance between two sequences is estimated as the evolutionary distance of the matrix that is optimal for their alignment. The optimal matrix can be found either by an iterative search for the ML matrix (Felsenstein 1989) or by integration to find the expected distance (EXP) (Agarwal and States 1996). Several matrix series exist that have been derived using different data and evolutionary models (Dayhoff, Schwartz, and Orcutt 1978; Jones, Taylor, and Thornton 1992; Müller and Vingron 2000; Whelan and Goldman 2001).

When applying distance-based approaches, tree reconstruction is conducted in two separate steps. First, pairwise distances are estimated between all sequences. Tree building is then based on the obtained pairwise distances. Previous studies have been limited because they only compared the performance of tree construction or distance methods in isolation (Russo, Naoko, and Nei 1996; Desper and Gascuel 2004). However, the best distance measure with one-tree method may not be the optimal choice for another tree algorithm. As real data always need to pass both distance estimation and tree reconstruction steps, it makes sense to evaluate the combined result of both steps. Another limitation with previous comparative studies is that they were done on only a handful of families.

In this paper, we present a large-scale evaluation of the most commonly used approaches and examine their ability to reconstruct the correct tree on a test set of 100 protein families. The performance of all 48 combinations of 12 distance measures and four tree construction methods was assessed. In order to further analyze the weaknesses and strengths of the methods, we stratified the results according to sequence divergence and distance from the tree leaves. All these methods have previously only been compared to a standard method (normally NJ) on small and different data sets. With this study, we are able to compare the performance of the different methods on the same data set, which due to its size and origin allows us to draw generalizable conclusions.

Methods

Standard NJ trees were constructed with Belvu 2.26. BIONJ as published, FastME as of February 28, 2003, and Weighbor 1.2.1 were used for the other methods. All methods were run with the default parameters suggested by the respective authors.

Pairwise protein distances were estimated from the alignment using lapd 1.0 (Arvestad 2004) for ML and EXP and Belvu 2.26 for standard Scoredist with Dayhoff calibration, JCP, Kimura, and uncorrected distance. For EXP and ML distances, the Dayhoff (Dayhoff, Schwartz, and Orcutt 1978), Jones-Taylor-Thornton (JTT) (Jones, Taylor, and Thornton 1992), Whelan and Goldman (WAG) (Whelan and Goldman 2001), and Müller-Vingron (MV) (Müller and Vingron 2000) matrices were used. All these methods ignore gaps.

The test set used to evaluate the tree construction and distance estimation methods was produced by generating multiple alignments from known trees. As the objective was to draw conclusions applicable to real life, trees were taken from the Pfam protein domain database (Bateman et al. 2004). From over 7,000 families, 100 families were selected with the requirement that they should contain 50–100 sequences and range between 100 and 500 columns in length. For the selected 100 families, guide trees for the simulated evolution were generated with the standard NJ algorithm as implemented in Belvu and uncorrected distance. An arbitrary member sequence was selected as root sequence to seed the simulated evolution process. Simulations are always a potential risk and can be questioned. They are, however, the only approach that gives verified tree/alignment combinations here, as the true tree for a given multiple sequence alignment is not known.

Each guide tree accompanied by the chosen root sequence was provided to ROSE 1.2 (Stoye, Evers, and Meyer 1998), which generated multiple alignments by simulated evolution. ROSE was run with three different evolutionary models on each family. This was done to investigate potential bias stemming from using a particular evolutionary model. By default, ROSE uses the Dayhoff transition probability matrix (Dayhoff, Schwartz, and Orcutt 1978). We also used the matrix described by Whelan and Goldman (2001). The third model was using the Dayhoff matrix for substitutions but disallowing deletion and insertion events (indels) by setting the indel probability to zero. Default ROSE parameters for version 1.2 were used for indel probabilities in the two data sets that allowed them, setting the probability of insertion or deletion events to 0.01. The lengths of indels were chosen according to a length function; here, also the default model was applied. ROSE can also be configured to use site-specific mutation rates, but this was not chosen in this study. For each model,

ROSE was run in tree replicates. In total, $100 \times 3 \times 3 = 900$ trees were generated for the test set. However, three trees failed on some methods, giving an effective size of $297 \times 3 = 891$ trees.

To assess the accuracy of the various tree-building methods and distance matrices, the reconstructed trees were compared to the original guide trees that were used by ROSE to generate the multiple sequence alignments. The comparison was made by calculating a "topological closeness measure" between the trees. A well-established topological distance measure was described by Robinson and Foulds (1981). A variant that is simpler to compute is available (Penny and Hendy 1985). The rationale is to consider every tree branch as a bipartition to all leaves. Comparing all bipartitions between two trees measures the topological difference. According to the original definition, the number of unmatched bipartitions is multiplied by 2 to yield the topological distance. Thus, the Robinson-Foulds topological distance between two bifurcating unrooted trees with n leaves is in the interval 0 (for isomorphic trees) and $2(n - 3)$ (if all leaves are placed differently). Trivial bipartitions of only single leaves are not considered. The Robinson-Foulds distance measure equals the necessary number of elementary operations (merging and splitting of nodes) to transform one tree into the other. The 297 used guide trees in the data set contained a total of 19,374 nontrivial bipartitions.

For this study, the accuracy of tree reconstruction was calculated as the inverse topological distance normalized to a score in the interval [0, 100], where 100 was attributed to identical trees and 0 for trees sharing no nontrivial bipartition. This choice is motivated, as each tree should contribute equally to the final result. The topological distance calculation makes no distinction between rooted and unrooted trees.

Results

A large-scale data set of multiple alignments corresponding to 100 trees with known topology was generated with the ROSE program. Simulation of multiple alignments using a guide tree and a model for amino acid substitutions, as in ROSE, is the only way to be absolutely sure about the true tree. The trees with known topology were derived from 100 medium-sized Pfam families. The pairwise identity among the sequences within one tree ranged between 99.7% and 26.1% with an average of 52.2%. They were generated by the standard NJ method, but in principle it does not matter how they were generated as the data used to build the trees were discarded and not in any way used for creating the test set alignments. This ensures that the analysis is not circular. A potential drawback of the simulation technique is that the generated sequences depend on the parameters in the evolutionary model. We therefore used three different models in ROSE to generate sequences from the guide trees: Dayhoff, WAG, and Dayhoff-nogap.

Figure 1 shows the overall result for each combination of the four tree-building methods and the 12 distance measures on data generated with the Dayhoff matrix. The standard "classical" implementation of the NJ algorithm is here carried out by Belvu (Sonnhammer 2005). BIONJ and Weighbor are recent improvements to the standard NJ algorithm, while FastME uses heuristics to find the ME tree.

The best result was obtained with BIONJ/Scoredist, while the worst result was produced by Weighbor/Kimura, with a difference in accuracy (topological closeness to the correct tree) of 4.5%. For a given distance measure, however, relatively small differences are observed in the accuracies of the four tree-building methods, generally less than 1% between the best method and the worst. This is expected from a large test set based on real trees that contain many "easy" and many "hard" nodes. The easy nodes will be found correctly by all methods, while the hard nodes will be found at low rates by different methods, and these differences tend to even out for many trees.

A clear trend is that BIONJ and Weighbor are more accurate than BelvuNJ and FastME. This is true for all distance measures except JCP and Kimura, where Weighbor becomes less accurate. BIONJ is the champion of tree building, winning with all distance measures except one (JTT ML) where it is marginally surpassed by Weighbor. On the other hand, Weighbor is mostly in second place only marginally behind BIONJ. UPGMA, an early and relatively primitive tree-building method, was also assessed on a subset of this study. As expected, the results were poor compared to those of BIONJ.

When comparing distance measures, the simple methods, uncorrected distance, JCP, and Kimura, were clearly outperformed by Scoredist, ML, and EXP. However, these distance measures give a fairly uniform impression. Scoredist reaches the highest accuracy for BIONJ and Weighbor, but several other measures come very close. For BelvuNJ, which is "standard NJ", Dayhoff ML was the best distance measure. It is interesting to note that JCP and Kimura were even inferior to uncorrected distance.

To investigate the results' dependence on the evolutionary model used in ROSE, we generated a test set from the same trees using the WAG (Whelan and Goldman 2001) transition probability matrix. Overall, the WAG results were similar to the Dayhoff results, therefore figure 1*B* shows the accuracy difference between WAG and Dayhoff for each method combination. As expected, distance estimation using the Dayhoff matrix was degraded on the WAG data set. We expected the WAG-based distances to improve correspondingly, but this was not the case except with FastME. However, FastME seems somehow to be heavily biased toward the WAG model as even the Dayhoff-based distance estimation gave better results on the WAG than on the Dayhoff test set. This increase was only minor, and because it was below the average increase for all methods it is shown with a negative value in the figure. FastME thus seems more robust and compatible with the WAG model than the other methods. The simple correction-based distance measures were also more accurate on WAG than on the Dayhoff data set.

We also investigated the dependence on the gap, or "indel", probabilities by generating a third version of the data set with the Dayhoff matrix and zero probability for indels, see figure 1*C*. In general, zero indel probabilities (which results in gapless alignments) should give higher

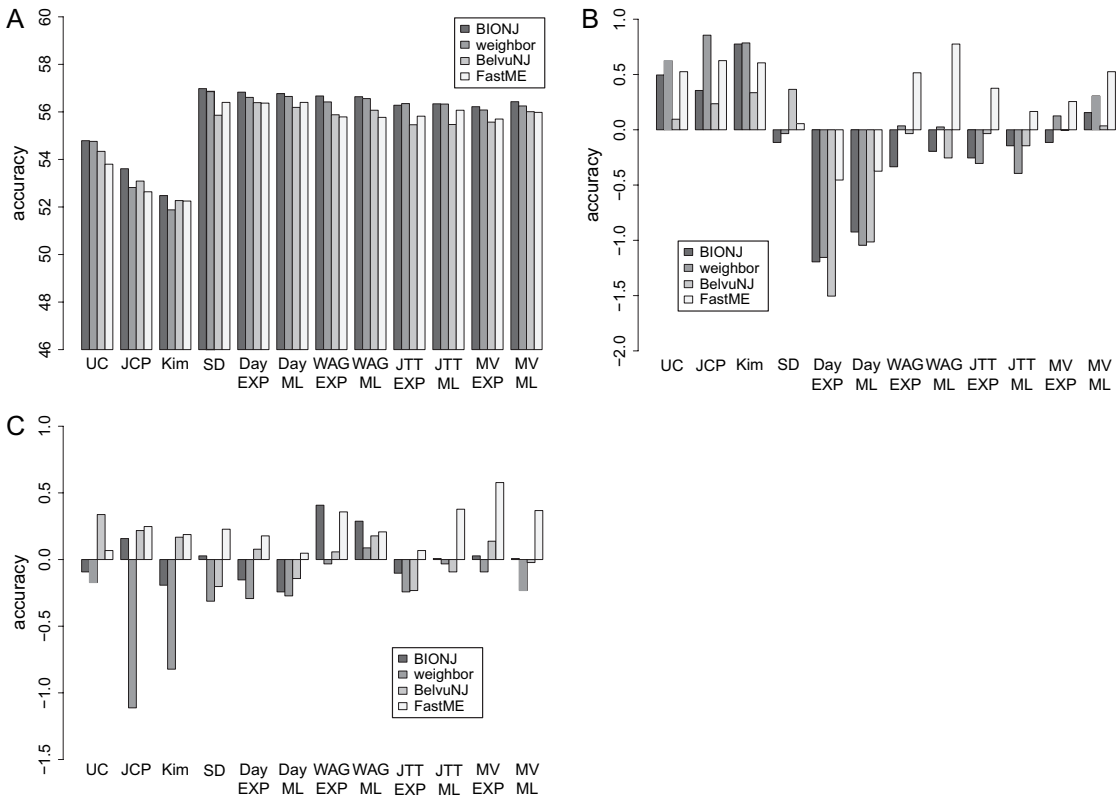

FIG. 1.—Evaluation of tree-building and distance estimation method combinations. The average accuracy obtained by combining four tree-building methods and 12 distance measures was evaluated on 100 trees of known topology. The accuracy was measured as topological closeness (see *Methods*) to the correct tree. For each model, 19,374 bipartitions were examined. (*A*) Data set generated with ROSE using the Dayhoff transition probability matrix. (*B*) The relative change in accuracy when instead generating the test set with the WAG matrix. Each value is the accuracy on WAG subtracted by the accuracy on Dayhoff −0.69, which was the average accuracy increase. This assigns positive values to methods that profited more than the average from the model change. Negative values down to −0.69 are attributed to methods with only minor gains, and values below −0.69 are given to methods which showed lower accuracy for the WAG model. (*C*) The relative change in accuracy contributed by indels in the data generation by ROSE with the Dayhoff matrix. The plot shows differences between the Dayhoff and the Dayhoff-nogap data set. Each value is accuracy with default indel probabilities subtracted by the accuracy with indel probabilities set to zero plus 1.27. The values are to be interpreted as in (*B*); however, the model change to indels gave an average accuracy decrease. Accuracy was measured as the fraction of nontrivial bipartitions shared between the true tree and the reconstructed tree. The evaluated distance estimators are uncorrected distance (UC), JCP, Kimura (Kim), Scoredist (SD), EXP, and ML distance. For EXP and ML distances, the JTT, Dayhoff (Day), WAG, and MV matrices were used. For each distance estimator, the tree-building methods are given in the same order: BIONJ, Weighbor, BelvuNJ, and FastME.

accuracies as more information can be used by the tree-building program. This was indeed observed and accuracy decreased by 1.27 on average when introducing indels. Weighbor was the tree-building method that suffered most from the model change. In fact, Weighbor was the most degraded method for 11 of the 12 distance measures. In contrast, FastME seems only lightly affected by the introduction of indels and was the least degraded method in 9 of the 12 cases.

To gain better insight into how the algorithms' performance depends on the tree-topological circumstance, we stratified the test set in two ways. Every branch was placed in a category based on its length, or based on the average distance between a branch and the leaves. The accuracy was again measured as topological closeness to the correct tree as described above. In principle, short branches should be hard to reconstruct and "interior" branches should also be more error prone.

The tree reconstruction accuracy was found to sharply increase with larger branch lengths, see figure 2. The increase was rather uniform for all tree-building methods using the Scoredist distance measure. However, for long branches FastME performed relatively worse than other algorithms. This was detected in all data sets (data not shown). For shorter branches, FastME was equally good as BIONJ and Weighbor; instead BelvuNJ performed slightly worse than the others here.

When comparing distance measures in the same way, a greater difference between methods was observed for short and intermediate branch lengths. The previous observation that uncorrected distance, JCP, and Kimura are less accurate than the rest is here more pronounced. It is of concern that of these three, uncorrected distance is the most accurate. The Scoredist method, although about as simple to implement as JCP or Kimura, was about as accurate as the more complex methods.

Figure 3 shows that tree accuracy also depended strongly on the distance to the leaves. All the tree-building methods (using Scoredist) were affected rather uniformly. The picture for different distance measures (using BIONJ) again showed that uncorrected distance, JCP, and Kimura performed substantially worse than the

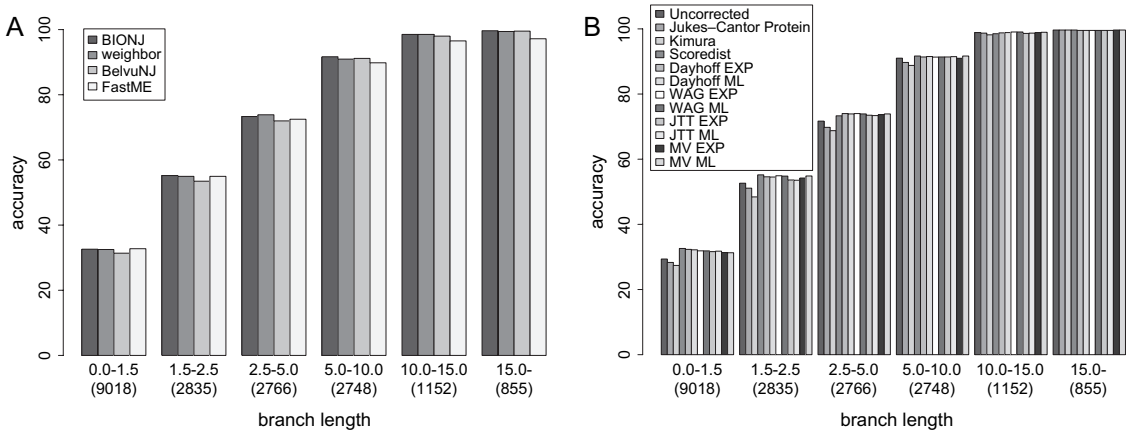

FIG. 2.—The ability to connect the correct tree branches was studied as a function of branch length (measured in PAM). (*A*) Comparing different tree reconstruction algorithms, all using the Scoredist distance estimator. (*B*) Comparing different distance measures, all used by the BIONJ tree algorithm. Shorter branch lengths were more difficult to reconstruct correctly for all methods, but some methods were degraded more than others. The number of evaluated bipartitions is noted within brackets.

other measures, uncorrected distance being the best of the three.

So far, only the average accuracy for a method combination over all families in the data set has been presented. However, the individual accuracy of different families varies between 35% and 80%. In order to analyze why some trees are easy and some hard to reconstruct, we compared the tree reconstruction accuracy to the variance in the repeated measurements of the same family, using the same data set and method (fig. 4). The variances were averaged for each family. There is a trend that high-accuracy trees have lower variance, but it is weak and hardly discernible up to 70% accuracy. However, the most accurate trees (>70% accuracy) showed significantly less variance.

Another reason for the variation in accuracy between families could be different levels of conservation. We examined this effect by calculating each family's average conservation with Belvu. The conservation of a column is the sum of all residue pair scores, according to BLOSUM62, divided by the total number of pairs. The average conservation of a family is the average over all columns. The conservation average from the three alignments per family was taken and plotted against the average accuracy for BIONJ/Scoredist (fig. 5). Plots for other methods than BIONJ/Scoredist had the same behavior (data not shown). Accuracy and conservation proved to be highly correlated, i.e., highly conserved families generally yield more accurately estimated trees. This trend is much stronger than the dependence of accuracy on the variance. Thus, in order to predict the confidence of a tree reconstruction, the main parameter is the level of conservation, while the variance between multiple measurements is less informative.

Phylogenetic tree construction is frequently applied to large data sets. Therefore, a suitable algorithm should have low complexity and good scalability. In this study, no tree contained more than 100 sequences. However, notable differences in runtime were already observed with this amount of data, see figure 6. Of the four algorithms evaluated in this study, Weighbor is by far the slowest. This has also been recognized by other authors (Desper and Gascuel 2002), who therefore did not use Weighbor with more than 100 sequences. The remaining methods

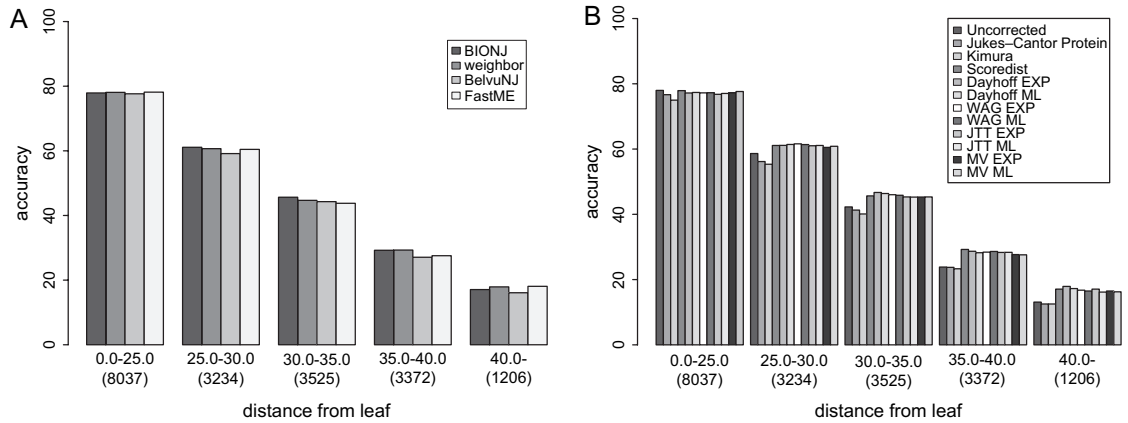

FIG. 3.—The ability to connect the correct tree branches was studied as a function of their average distance from the leaves. (*A*) Comparing different tree reconstruction algorithms, all using the Scoredist distance estimator. (*B*) Comparing different distance measures, all used by the BIONJ tree algorithm. The more distant a branch is from the leaves, the more difficult it was for all methods to reconstruct, but some methods were degraded more than others.

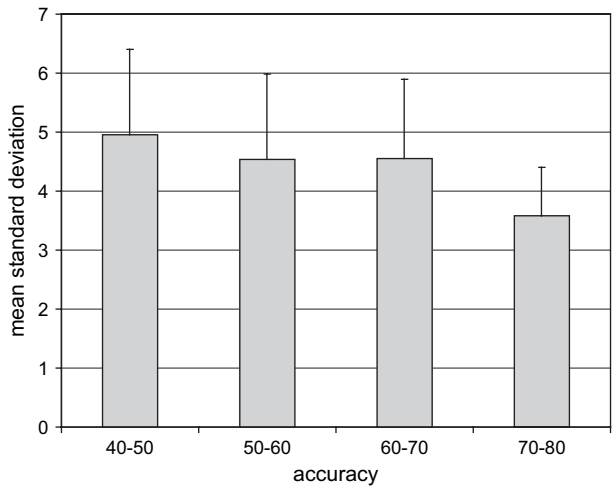

FIG. 4.—Correlation between tree accuracy and its variation. The mean accuracy (topological closeness to correct tree) was calculated as the average for each family over all data sets, distance methods, and replicates ($3 \times 12 \times 3 = 108$ measures for each family) using the BIONJ tree-building method. The standard deviation was calculated using the replicates with the same parameters, and the mean was taken over all data sets and distance methods ($3 \times 12 = 36$ measures for each family). There is a trend for lower variance in higher accuracy trees. Only five families had a mean accuracy below 40% and were omitted.

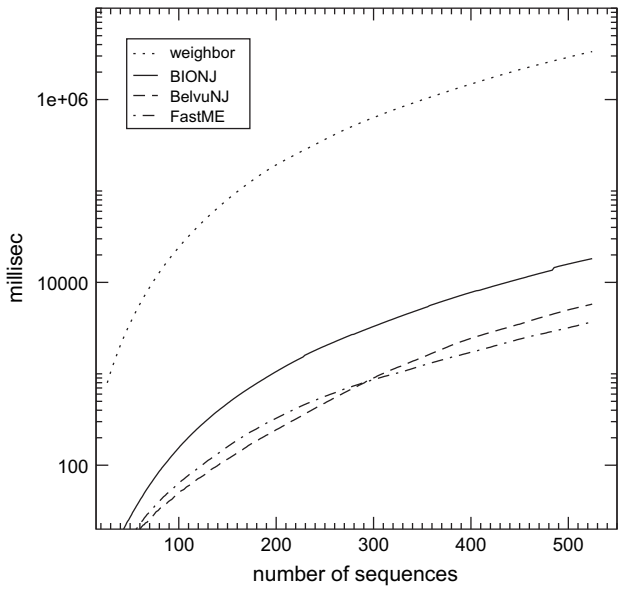

FIG. 6.—The tree construction algorithms show large variances in time consumption. Times are noted as user time on our system. Weighbor is by far slowest and may be used only for a limited number of sequences. Among the other methods, BelvuNJ is fastest up to 300 sequences. For a larger number of sequences, FastME takes the lead. Both of them are always faster than BIONJ.

did not differ much in computational speed. However, FastME seemed faster than BIONJ, which has also been noted before (Desper and Gascuel 2004).

The guide trees for simulation were calculated by NJ from manually curated Pfam alignments. This raises the question whether this is the reason why NJ-based algorithms performed best in the study. To investigate this, we also generated a data set using UPGMA guide trees from the same Pfam alignments and made simulated multiple alignments the same way as before. The result of this was that although the UPGMA-reconstructed trees obtained increased accuracy and NJ-reconstructed trees obtained decreased accuracy, the NJ trees were still superior. Using the Dayhoff data model allowing for indels for simulation and comparing BIONJ/Scoredist with UPGMA/Scoredist, the accuracy was increased by 3.4% points for UPGMA and decreased by 4.3% points for BIONJ. However, BIONJ was still 3.5% points more accurate than UPGMA, thus supporting the generality of the overall results in this study.

To allow researchers evaluating their own methods, we have made our data publicly available at ftp://ftp.cgb.ki.se/pub/data/treedist/dataset.tgz.

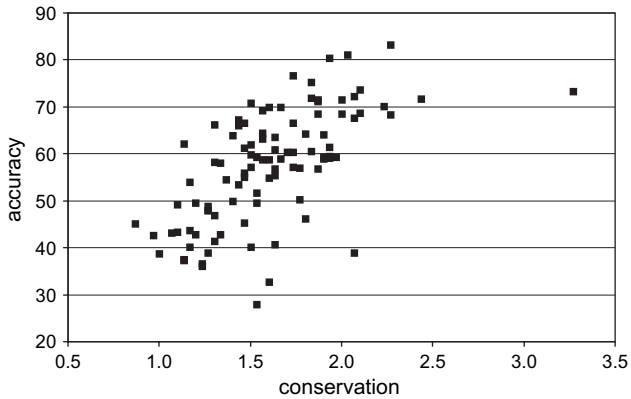

FIG. 5.—Correlation between family conservation and tree accuracy. The average conservation was computed by Belvu for all three alignments per family and were plotted against the average accuracy of the estimated trees. For simplicity only the results for the most accurate method combination, BIONJ/Scoredist, are included.

Discussion

This study is unique in two ways: it evaluates a large number of distance-based tree-building methods on a large test set. The quality of the distance estimates is not measured directly but only as the correctness of reconstructed tree topologies. This is motivated by the fact that most users are more interested in getting the topology of the tree correct than the lengths of the branches.

The conclusion from the comparison of distance estimation algorithms was that there is essentially no difference between the "optimal matrix" methods and Scoredist, while the old correction methods Kimura and JCP performed significantly worse, even worse than uncorrected distance.

The conclusion from the comparison of tree-building algorithms was that the enhancements to the original NJ algorithm improve the accuracy to some extent, but rather marginally, typically less than 1%. The measurement of improvement is of course dependent on the data set, but as this was constructed from phylogenetic trees made from real alignments the observed accuracies are realistic.

In terms of speed, all algorithms except Weighbor are roughly equal. Weighbor was clocked about 200 times slower than the rest, which reduces its usefulness. This high

time consumption does not seem to be motivated as Weighbor was generally less accurate than the much faster BIONJ. Compared to other tree-building approaches, e.g., ML or Bayesian statistics (Ronquist and Huelsenbeck 2003) that do not employ pairwise distances, Weighbor is not a slow method. Likelihood methods can still only be applied to small data sets. Williams and Moret (2003) report that some of these methods cannot handle alignments of 40 sequences, which was below the minimum sequence number in this study. Likelihood methods have therefore not been included here.

The authors of FastME have repeatedly claimed higher accuracy of their method (Desper and Gascuel 2002, 2004). Our findings did not confirm this observation and are in fact more in line with results of Bruno (2004). Bruno reported a long-branch attraction bias for FastME. This enables FastME to successfully work on data with an underlying molecular clock, but when this is violated FastME performs worse, particularly on long branches.

Previous studies concluded that the quality of the underlying multiple alignment may play a greater role for the accuracy than the choice of tree reconstruction method (Morrison and Ellis 1997). Because the multiple alignments were obtained directly from ROSE, the correct multiple alignment was always known in our study. As the multiple alignments were not constructed by a multiple alignment program, no particular bias to one method or the other should be present.

Acknowledgment

This work was supported by a grant from Pfizer Corporation.

Literature Cited

Agarwal, P., and D. J. States. 1996. A Bayesian evolutionary distance for parametrically aligned sequences. J. Comp. Biol. **3**:1–17.

Arvestad, L. 2004. Estimating protein distances. http://www.nada.kth.se/~arve/lapd.

Atteson, K. 1997. The performance of the NJ method of phylogeny reconstruction. Pp. 133–147 *in* B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds. Mathematical hierarchies and biology, DIMACS series in discrete mathematics and theoretical computer science, Vol. 37. American Mathematical Society, Providence, R. I.

Bateman, A., L. Coin, R. Durbin et al. (10 co-authors). 2004. The Pfam protein families database. Nucleic Acids Res. **32**(Database Issue):D138–D141.

Bruno, W. J. 2004. How does FastME compare with Weighbor? http://www.t10.lanl.gov/billb/weighbor/fastme/.

Bruno, W. J., N. D. Socci, and A. L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. **17**:189–197.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed. Atlas of protein sequence and structure, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

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

———. 2004. Theoretical foundation of the balanced minimal evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. **21**:587–598.

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

———. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics **5**:164–166.

Gascuel, O. 1997. BIONJ: an improved version on the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

Howe, K., A. Bateman, and R. Durbin. 2002. QuickTree: building huge Neighbour-Joining trees on protein sequences. Bioinformatics **18**:1546–1547.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* Munro H. N., ed. Mammalian protein metabolism. Academic Press, New York.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Mailund, T. 2003. QuickJoin. (http://www.daimi.au.dk/~mailund/quick-join.html).

Morrison, D. A., and J. T. Ellis. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNA of apicomplexa. Mol. Biol. Evol. **14**:428–441.

Müller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comput. Biol. **7**:761–776.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. **30**:371–403.

Penny, D., and M. D. Hendy. 1985. The use of tree comparison metrics. Syst. Zool. **34**:555–566.

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

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574.

Russo, C. A. M., T. Naoko, and M. Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. **13**:525–536.

Saitou, N., and T. Imanishi. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514–525.

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

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. Freeman, San Francisco, Calif.

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. **28**:1409–1438.

Sonnhammer, E. L. L. 2005. Belvu. (ftp://ftp.cgb.ki.se/pub/prog/belvu/).

Sonnhammer, E. L. L, and V. Hollich. 2005. Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics **6**:108.

Stoye, J., D. Evers, and F. Meyer. 1998. Rose: generating sequence families. Bioinformatics **14**:157–163.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12**:823–833.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**:691–699.

Whelan, S., P. Lio, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. **17**:262–272.

Williams, T. L., and B. M. E. Moret. 2003. An investigation of phylogenetic likelihood methods. Pp. 79–86. Proceedings of the 3rd IEEE Symposium on Bioinformatics and Bioengineering. IEEE Press, Piscataway, N.J.

Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acids sequences inferred by parsimony, likelihood and distance methods. J. Mol. Evol. **44**:139–146.

Manolo Gouy, Associate Editor

Accepted July 22, 2005