

Exploring the Relationship between Sequence Similarity and Accurate Phylogenetic Trees

Brandi L. Cantarel,* Hilary G. Morrison,† and William Pearson*

*Department of Biochemistry and Molecular Genetics, University of Virginia and †Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts

We have characterized the relationship between accurate phylogenetic reconstruction and sequence similarity, testing whether high levels of sequence similarity can consistently produce accurate evolutionary trees. We generated protein families with known phylogenies using a modified version of the PAML/EVOLVER program that produces insertions and deletions as well as substitutions. Protein families were evolved over a range of 100–400 point accepted mutations; at these distances 63% of the families shared significant sequence similarity. Protein families were evolved using balanced and unbalanced trees, with ancient or recent radiations. In families sharing statistically significant similarity, about 60% of multiple sequence alignments were 95% identical to true alignments. To compare recovered topologies with true topologies, we used a score that reflects the fraction of clades that were correctly clustered. As expected, the accuracy of the phylogenies was greatest in the least divergent families. About 88% of phylogenies clustered over 80% of clades in families that shared significant sequence similarity, using Bayesian, parsimony, distance, and maximum likelihood methods. However, for protein families with short ancient branches (ancient radiation), only 30% of the most divergent (but statistically significant) families produced accurate phylogenies, and only about 70% of the second most highly conserved families, with median expectation values better than 10^{-60} , produced accurate trees. These values represent upper bounds on expected tree accuracy for sequences with a simple divergence history; proteins from 700 *Giardia* families, with a similar range of sequence similarities but considerably more gaps, produced much less accurate trees. For our simulated insertions and deletions, correct multiple sequence alignments did not perform much better than those produced by T-COFFEE, and including sequences with expressed sequence tag-like sequencing errors did not significantly decrease phylogenetic accuracy. In general, although less-divergent sequence families produce more accurate trees, the likelihood of estimating an accurate tree is most dependent on whether radiation in the family was ancient or recent. Accuracy can be improved by combining genes from the same organism when creating species trees or by selecting protein families with the best bootstrap values in comprehensive studies.

Introduction

Despite the expectation that abundant genome and expressed sequence tag (EST) sequence data would produce better evolutionary histories, particularly for organisms from deep branches in the “tree of life,” phylogenies based on genome-scale data are often inconsistent (de la Cruz and Davies 2000; Eisen 2000; Ochman et al. 2000; Jain et al. 2003). One explanation for inconsistent bacterial phylogenies is a high frequency of lateral gene transfer (LGT) (Doolittle 1999; Mirkin et al. 2003), which has led some evolutionary biologists to question whether life can be meaningfully represented as a tree. LGT is recognized as a major evolutionary mechanism (Doolittle 1998; Lopez-Garcia and Moreira 1999; Campbell 2000), which is inferred when a homolog from a distantly related species is more similar than homologs from more closely related species or when trees from different gene families in the same organisms are inconsistent. True instances of LGT can be quite difficult to distinguish from artifacts of computational methods, such as long-branch attraction (LBA), biased taxon sampling, or gene loss (Eisen 2000). Problems caused by biased taxon sampling should be reduced as more genomes are sequenced.

Compared with LGT, computational artifacts such as LBA provide a less provocative explanation for inconsistent phylogenies. LBA can result in the grouping of the fastest evolving taxa, irrespective of their true phylogenies (Felsenstein 1978), and is expected for sequences from

deeply branching or rapidly evolving organisms. LBA can be reduced by increasing or modifying the taxonomic sample (Hillis 1996; Graybeal 1998; Hillis 1998; Zwickl and Hillis 2002), by focusing on slowly evolving positions (Felsenstein 1978; Olsen 1987; Brinkmann and Philippe 1999), and by improving the model of sequence evolution (Lockhart et al. 1996; Yang 1996; Hirt et al. 1999; Van de Peer et al. 2000). In general, LBA is thought to play a less important role than LGT when large numbers of proteins, some of which share strong sequence similarity, are considered.

It has been difficult to estimate the relative importance of LGT (biological) and LBA (methodological) as explanations for inconsistent trees, in part because we do not know how often phylogenetic methods fail for sequences that evolved vertically, and that can be identified by widely used sequence comparison methods like Blast (Altschul et al. 1990) and FASTA (Pearson and Lipman 1988). In this paper, we explore the relationship between sequence similarity and phylogenetic accuracy from various tree topologies and evolutionary models using protein sequences that share statistically significant similarity (i.e., that could be identified in a similarity search). We created simulated gene families over a broad range of evolutionary distances, using simple models of rate variation and gap insertion, and measured the accuracy of phylogenetic estimation methods. We were surprised to find that despite our simple evolutionary models, 10–40% of the reconstructed trees were less than 80% accurate, even for very highly conserved simulated protein families.

Methods

Generating Protein Families

Protein families were created using a modified version of EVOLVER (Yang 1997; Yang et al. 1998), a program

Key words: simulation, phylogenetic analysis, accuracy, sequence similarity.

E-mail: wrp@virginia.edu.

Mol. Biol. Evol. 23(11):2090–2100. 2006

doi:10.1093/molbev/msl080

Advance Access publication August 4, 2006

© 2006 The Authors

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

that evolves sequences along a specific tree. EVOLVER was modified to generate insertions and deletions using the Benner model (Benner et al. 1993). We used the codon evolution model in EVOLVER, which requires a root sequence length, an input tree, and codon frequencies. Protein sequence lengths were chosen from a normal distribution with mean 468 ± 173 , derived from protein sequence lengths of 11 completely sequenced eukaryotic genomes. We sought a range of evolutionary distances comparable with those likely to be encountered when characterizing proteins from primitive eukaryotes, with point accepted mutation (PAM; Dayhoff et al. 1972, 1978) values ranging from 100 to 400, which correspond to EVOLVER total branch lengths from 2.85 to 11.4. The EVOLVER input branch lengths were determined empirically by producing protein families over a range of branch lengths, aligning the members of the family, and calculating the PAM distance of the most distant members. Codon frequencies were calculated from the mRNA sequences of 6 completely sequenced eukaryotic genomes (*Apis mellifera*, *A. thaliana*, *Encephalitozoon cuniculi*, *Plasmodium falciparum*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*). Otherwise, the default parameters were used: omega (dN/dS) = 0.3, kappa (transition/transversion) = 5, and a uniform distribution of rate variation along the codon sites. Because deep phylogenies are more accurately estimated with protein than with DNA (at distances greater than PAM50, the underlying DNA sequences would not share statistically significant sequence similarity), genes were translated into protein sequences (Hall 2005).

Assessing Evolutionary Distances

To characterize the diversity of each protein family, all its members were compared using PRSS (Pearson and Lipman 1988), which estimates the statistical significance of a sequence similarity by shuffling one of the sequences. The expectation value ($E(\cdot)$ value) calculated by PRSS is equivalent to the $E(\cdot)$ value calculated by Blast; it reports the number of times a similarity score is expected to occur by chance in a database search against a sequence database of specified size. Here, we report $E(10,000)$ —the number of times an alignment score would be found by chance in a search of a 10,000 entry database. The protein families in this study have either 8 (unbalanced trees) or 16 (balanced trees) members. For each of the 8 (or 16) members of a family, 7 (or 15) pairwise expectation values were calculated. The median expectation value for the family is the lowest (most significant or most closely related) of the 8 (or 16) median expectation values of the individual family members.

Multiple Sequence Alignment Accuracy Analysis

Protein families were aligned using T-COFFEE (Notredame et al. 2000). True multiple sequence alignments were created using the known positions of changes at homologous sites produced by EVOLVER. Multiple alignment accuracy was evaluated with the VerAlign comparison software, which is available at www.ibivu.cs.vu.nl/programs/veralignwww/ (Simossis et al. 2005). VerAlign calculates a sum-of-pairs (SOP) score that reports the number of correctly aligned residues in each of the pair-

wise residue alignments implied by the multiple sequence alignment.

Phylogenetic Trees of Synthetic Families

Phylogenies were constructed using 4 methods: 1) distance (PHYLIP—Fitch, Fitch and Margoliash 1967; Felsenstein 1993), 2) maximum parsimony (PHYLIP—Protpars, Eck and Dayhoff 1966; Felsenstein 1993), 3) maximum likelihood (PHYML, Guindon and Gascuel 2003), and 4) Bayesian inference (MrBayes, Ronquist and Huelsenbeck 2003). For distance trees, each multiple sequence alignment of a protein family was used to make a distance matrix using PROTDIST (Felsenstein 1993), with the Jones–Taylor–Thornton method of amino acid replacement (Jones et al. 1992). Maximum likelihood (PHYML) reconstruction used Jones–Taylor–Thornton protein rate matrices and included the default gamma parameter for an “average” distribution of evolutionary rate variation over the protein sites (Guindon and Gascuel 2003). Bayesian analysis used PAM250 rate matrices (Dayhoff et al. 1978) and was done with and without including gamma-distributed rate variation across the sites. Bootstrap analysis (Felsenstein 1985) was performed using maximum likelihood.

Tree Evaluation

Tree accuracy was evaluated using 2 programs: TREEDIST (Felsenstein 1993) and TREESCORE (this work). TREEDIST calculates 2 measures of tree accuracy: symmetric difference (Robinson and Foulds 1981) and branch score difference (Kuhner and Felsenstein 1994). Symmetric difference measures the differences in the topology of 2 trees, whereas the branch score difference measures differences in topology and branch length as a root sum of squares. We normalized both distances by dividing them by the corresponding maximum distances of trees with random topologies.

Similar to CompareTree (Hall 2005), we also developed a measure of tree accuracy that directly counts the correct number of clades—TREESCORE. Each ancestor node and its descendents in the model tree represent a clade. The TREESCORE is the ratio of the number of correct clades contained in the estimated tree divided by the maximum possible number of clades. Clades containing only 2 leaf sequences were weighted half, to focus the measure on deeper clades that are more difficult to reconstruct. Thus, for a balanced unrooted tree (fig. 1A) with 16 taxa, the maximum number of correct clades is 9. For the unbalanced unrooted tree in figure 1C, the maximum number of correct clades is 4. In the analysis of the *Giardia* families, the “correct” clades are deep, so leaf clades were not downweighted.

Giardia Phylogenies

For purposes of high-throughput phylogenetic tree generation for *Giardia*, each open reading frame (Morrison et al. 2004) was individually searched against a custom reference protein database using BlastP, with the BLOSUM62 substitution matrix, and default parameters (Altschul et al. 1990). For purposes of multiple sequence alignment, a single putative ortholog was retained from each reference genome

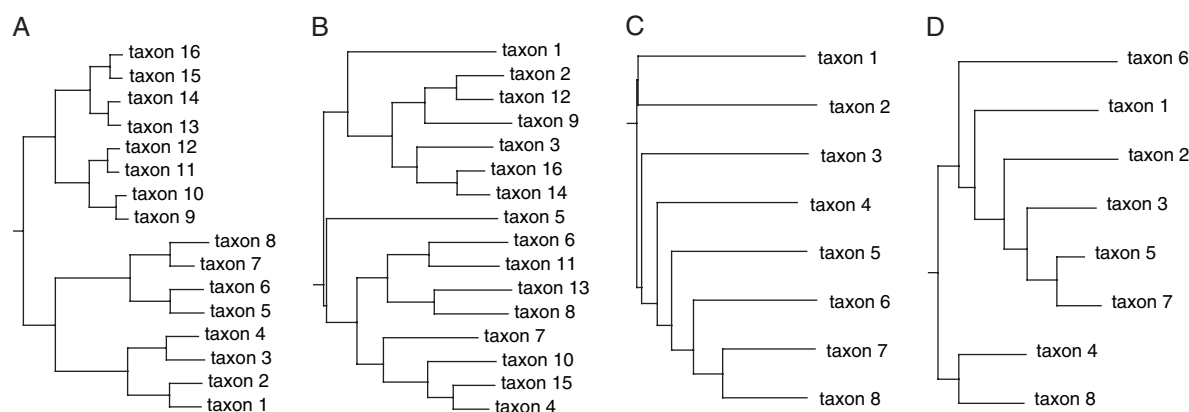


FIG. 1.—Model tree topologies. EVOLVER trees produced from the 4 topologies used in this study. (A, B) Balanced tree topologies and (C, D) asymmetrical tree topologies with (A) recent, (B, D) uniform, and (C) ancient radiation. The differences in branch lengths in each of the panels, but particularly (B), and (D), reflect the random variation in branch lengths produced by the EVOLVER program.

if it had a Blast expectation value $\leq 10^{-10}$. Multiple alignments were performed with T-COFFEE. Phylogenies were estimated using Mr. Bayes (Ronquist and Huelsenbeck 2003) or PHYML (Guindon and Gascuel 2003). Trees were assessed with TREESCORE with the following clades: animals (*Caenorhabditis elegans* and *Mus musculus*), plants (*A. thaliana* and *Oryza sativa*), bacteria (*Aquifex aeolicus*, *Escherichia coli*, and *Rickettsia prowazekii*), apicomplexans (*Cryptosporidium parvum*, *P. falciparum*, and *Toxoplasma gondii*), kinetoplastids (*Leishmania major*, *Trypanosoma cruzi*, and *Trypanosoma brucei*), fungi (*S. cerevisiae* and *Cryptococcus neoformans*), stramenopiles (*Phytophthora sojae*, *Phytophthora ramorum*, and *Thalassiosira pseudonana*), archaea (*Archaeoglobus fulgidus* and *Sulfolobus solfataricus*), and plants/algae (*Oryza sativa*, *A. thaliana*, and *Chlamydomonas reinhardtii*).

Synthetic EST Sequences

To examine the effect of EST errors in phylogeny estimation, we added additional substitution and insertion/deletion errors to the DNA sequences produced by EVOLVER and truncated the sequences to produce simulated ESTs. EST error rates were estimated by comparing *Giardia lamblia* EST sequences with the *Giardia lamblia* Genome Database (Morrison et al. 2004) using FASTA (Pearson and Lipman 1988) with match/mismatch +1/−3 and gap penalties −12/−4. Alignments were analyzed for the number of mismatches, insertions, and deletions in the EST sequences compared with the genome sequence. EST sequences had about 2.4% substitutions per read with 0.075% insertions and 0.053% deletions. Assembled contigs had about one-third as many errors (0.72%, 0.032%, and 0.018%). DNA sequences from each protein family were mutated according to an EST-like or EST-contig-like model. Because truncations of EST sequences occur at the 5' end, sequence truncation was incorporated into the model, based on the EST truncations seen in *Giardia* EST sequences. Only 16% of ESTs aligned with the N terminus of the protein; 36% aligned over 75% of the protein's length; 56% aligned over half the protein; and 84% of the ESTs covered at least 25% of the protein-coding region. We also considered a "limited truncation model," in which the

EST started and ended randomly within the protein-coding region (based on the distribution of *Giardia* EST alignments), but ESTs were excluded if less than 50% of the protein was aligned. Protein translations were determined by comparing the sequence with other members in the family using FASTY (Pearson et al. 1997).

Sequence Concatenation

We generated 2 data sets of concatenated sequences. Proteins were concatenated from the same taxa evolved under the same tree topologies and model trees. In addition, we concatenated sequences that were evolved to the same evolutionary distance. These concatenated sequences were then analyzed as individual sequences.

Results

Because many phylogenomic studies rely on large-scale alignments of protein and DNA sequences that were identified by sequence similarity searches, we explored the relationship between sequence similarity and phylogenetic accuracy. We generated 2,400 protein sequence families from 48 model trees produced from 1) 4 tree topologies (fig. 1), 2) 3 radiation models, and 3) 4 evolutionary distances. The tree topologies consisted of two 16-taxon balanced (fig. 1A and B) and two 8-taxon unbalanced (fig. 1C and D) topologies; to preserve the distribution of internode distances, the unbalanced trees had fewer taxa. Branch lengths were specified using 3 radiation models: recent, uniform, and ancient. Recent radiation protein families have more evolutionary change before speciation events (fig. 1A), resulting in longer internal branches and shorter "leaf" branches. In recent radiation families, daughter branches were half the evolutionary distance of the parent branch, whereas in ancient radiation families (fig. 1C), daughter branches were twice as long as the parent. Ancient radiation trees are expected to be the most difficult to reconstruct accurately because of their short internal branches. The evolutionary distance of the internal branches provided to EVOLVER was constant for uniform radiation protein families, but EVOLVER introduces variation into the branch lengths of the trees it generates (fig. 1B). Even though there

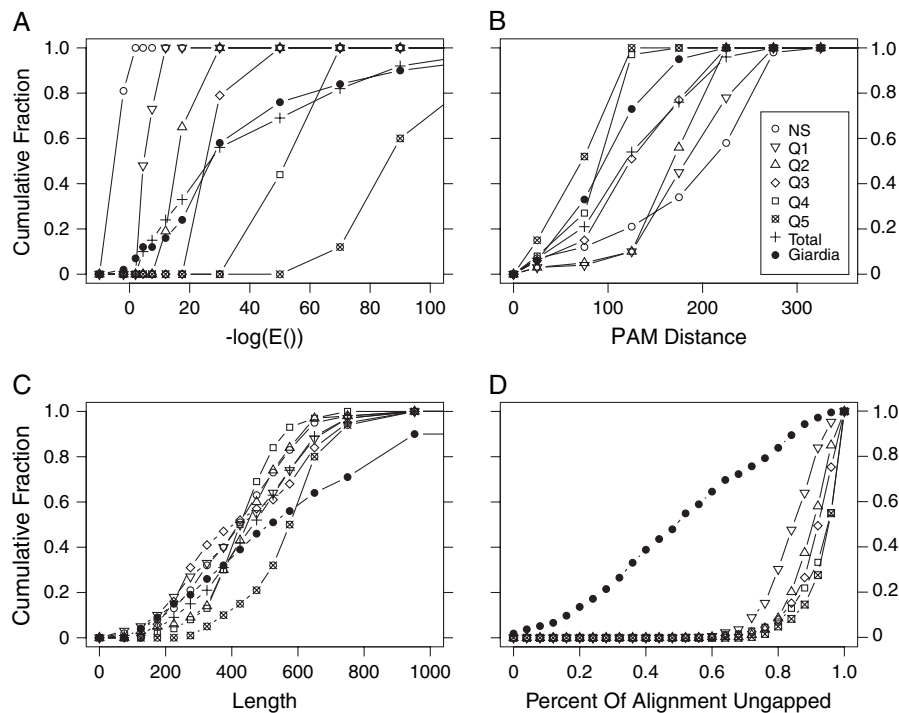


FIG. 2.—Ranges of sequence similarity and evolutionary distance. In all, 2,400 protein families, created using different divergence models and topologies, were divided into 6 groups based on median expectation value (diversity). Two measures of diversity are shown: (A) statistical significance in expectation value and (B) similarity in PAM distance. Panel (C) shows the distribution of family lengths in each group. (D) The median ungapped multiply aligned region divided by the median protein length, for each family. Sequences that do not share statistically significant median sequence similarity (NS) are represented by open circles ($\circ$). Quintiles of sequences that have statistically significant median similarity are indicated by least similar (Q1, ∇), second least similar (Q2, $\triangle$), third least similar (Q3, $\diamond$), fourth least similar (Q4, $\square$), and most similar (Q5, $\boxtimes$). The distribution for all simulated gene families that have significant median expectation values is shown (+), as are the *Giardia* families ($\bullet$) found in 6 of 9 taxon groups.

is a strong correlation between branch lengths provided to the EVOLVER program and evolutionary distances of the resulting trees, the rate variation used by EVOLVER can produce phylogenies with distributions of evolutionary distance that partially obscures the evolutionary model (fig. 1D).

To explore the relationship between measured sequence similarity and phylogenetic accuracy, target evolutionary distances for the most distant sequences in the family ranged from 100 to 400 PAMs (Dayhoff et al. 1972, 1978). Protein families were grouped by median expectation value from the most divergent to the least divergent. About 37% of sequences did not share statistically significant sequence similarity and therefore would not have been found with a Blast search (Altschul et al. 1990). The remaining families, which shared significant median sequence similarity, were split into quintiles, from least similar (Q1) to most similar (Q5). As expected, groups with the best statistical significance (fig. 2A) had the lowest sequence divergence (fig. 2B). All 6 groups have similar distributions of lengths (fig. 2C), with the exception of the least divergent quintile (Q5), which had longer sequences, on average. Because we grouped families by median statistical significance, which depends both on sequence similarity and sequence length, quintiles overlap in their ranges of PAM distance. In addition to the sequence variation produced by the EVOLVER program codon model, we introduced insertions and deletions into the protein sequences, using parameters estimated by Benner et al. (1993), based on structural alignments (fig. 2D).

To confirm that our simulated protein families span a range of evolutionary distances similar to those likely to be encountered in a challenging eukaryotic data set, we examined 705 *Giardia* proteins with statistically significant homologs in 6 of 9 groups of organisms. The distributions of statistical significance, evolutionary distance, length, and ungapped alignment lengths for the *Giardia* protein families are also shown in figure 2. The conserved *Giardia* protein families show a distribution of sequence similarity and evolutionary distance that is very similar to the overall distribution of similarity and distance of our synthetic families that share statistically significant similarity (fig. 2, comparing + total with $\bullet$ *Giardia*). However, the *Giardia* proteins have substantially shorter regions of ungapped sequence in the multiple alignments than the synthetic proteins do (fig. 2D). This presumably reflects domain insertions in *Giardia* and deletions, which are not part of the Benner insertion/deletion model.

The median similarity characterizes the “average” properties of each protein family; however, the ability of a family to reconstruct a phylogenetic tree accurately may depend more on the most distant members of the family. Table 1 summarizes the radiation models—ancient, uniform, or recent—in the different quintile groups and the statistical significance of the most distant members in each quintile. Thus, 72% of the proteins in the most distant quintile (Q1) were produced with either the ancient or the recent radiation model; the remaining families came from the uniform model. Surprisingly, 14% of the least divergent

Table 1
Characteristics of Protein Family Groups

Radiation Model		Fraction of Families from Radiation Model ^a			Fraction of Families with a Nonsignificant Alignment ^b		
Group	<i>E</i> (10,000)	Ancient	Uniform	Recent	Ancient	Uniform	Recent
Q1	1×10^{-6}	0.28	0.43	0.28	0.21	0.41	0.24
Q2	1×10^{-19}	0.33	0.39	0.29	0.01	0.11	0.04
Q3	1×10^{-31}	0.34	0.37	0.29	0.00	0.02	0.07
Q4	1×10^{-61}	0.29	0.31	0.40	0.00	0.00	0.14
Q5	1×10^{-94}	0.16	0.28	0.56	0.00	0.00	0.14

^a Fraction of families from each of the 3 radiation models. Each quintile contains 300 protein families.

^b Fraction of families with at least one pairwise alignment with a nonsignificant alignment ($E(10,000) > 10^{-3}$).

families produced with the recent radiation model have at least 1 pair of members that lacks statistically significant similarity ($E(\cdot) > 10^{-3}$). In practice, these family members could be identified through transitive homology or with a profile sequence similarity program like PSI-Blast (Altschul et al. 1997). More of the ancient radiation protein families fall in the not-significant category; as a result, in all the significant quintiles, there are higher percentages of families from uniform or recent radiation models.

Multiple Sequence Alignment Accuracy

Sequence-based phylogenetic reconstruction strategies require multiple sequence alignments; inaccurate multiple alignments are often proposed to explain inconsistent phylogenies. To assess the multiple sequence alignment accuracy at different evolutionary distances, we compared our estimated T-COFFEE alignments with the true multiple sequence alignments recorded as the families were produced by EVOLVER. Sequence families that did not share statistically significant sequence similarity, and thus would not have been detected in a Blast search, were poorly aligned with T-COFFEE (fig. 3). More than half of these estimated alignments had SOP scores that were 60% of the scores produced by true alignments. However, over 90% of families with statistically significant sequence similarity had SOP scores greater than 80% of the true alignment (fig. 3). As expected, the estimated alignments look more like the true multiple sequence alignments as the amount of divergence decreases.

Phylogenetic Accuracy

To determine the relationship between evolutionary distance and tree accuracy, we estimated phylogenies for our simulated families using 4 different methods: 1) distance (fig. 4A), 2) parsimony (fig. 4B), 3) maximum likelihood (fig. 4C), and 4) Bayesian (fig. 4D). Our TREESCORE threshold for accuracy was 0.80; phylogenies at that score have only 1 (for 8-taxon families) or 2 (for 16-taxon families) incorrectly grouped clades within the tree. More than 95% of the recent radiation families—families with longer ancient branches—were more than 80% accurate (fig. 4) over the entire range of sequence similarities, including families whose members did not share statistically significant similarity. For ancient radiation

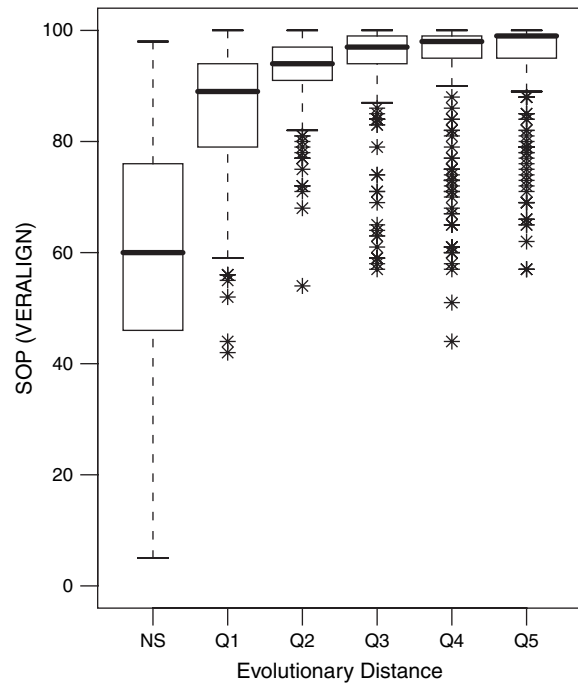


FIG. 3.—Multiple sequence alignment accuracy. Sequence families were aligned with T-COFFEE. The resulting alignments were compared with the true alignment with VERALIGN. SOP is the sum of pairs, which is number of columns that differ between the inferred and the correct alignment (regardless of whether they contain residues or gap characters), divided by the number of columns in the alignment, summed over all the pairwise alignments. Plots were drawn using the “R” box plot() function, which shows the median, 25% and 75% percentiles and draws whiskers at 1.5 times the interquartile (25–75%) range. The asterisks indicate outliers—individual families outside the whiskers.

families—families with short ancient branches—80% accuracy was achieved in only 15–50% of the most divergent families that shared statistically significant similarity, depending on the phylogenetic reconstruction method used. Sixty to 90% of families from the uniform radiation model produced 80% accurate trees from the most divergent, but statistically significant, quintile (fig. 4C). All 4 tree-building methods produced similar trends, with each method performing better with some radiation models at some distances.

We measured phylogenetic accuracy by 3 ways: symmetric tree distance (Robinson and Foulds 1981), branch score difference tree distance (Kuhner and Felsenstein 1994), and TREESCORE. Figure 5 shows the results for maximum likelihood trees, but similar results are found with distance, parsimony, and Bayesian phylogenies as well (data not shown). For each measure of accuracy, more closely related protein families produce more accurate trees. However, it can be difficult to correlate TREEDIST distances to the number of correct clades. Because biologists are interested in the number of clades, we evaluated a metric that measures this directly. TREESCORE gives a distribution of scores (fig. 5C) similar to the symmetric tree distances calculated by TREEDIST (fig. 5A), and those scores increase as the diversity of the protein families decreases.

To evaluate the ability of our simulated families to reproduce the phylogenetic accuracy of actual protein

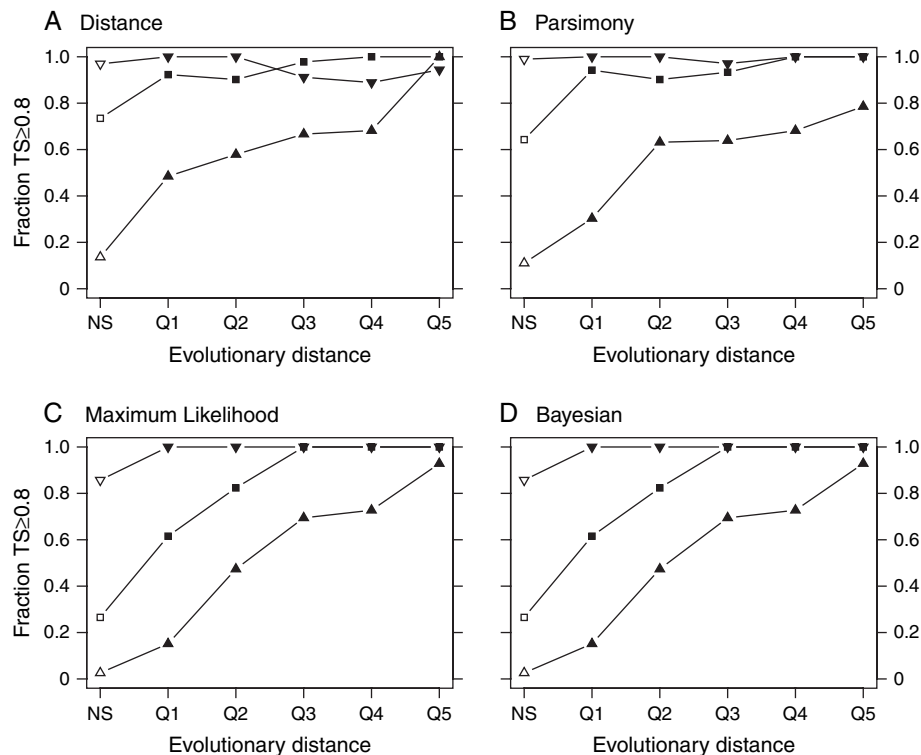


FIG. 4.—Phylogenetic accuracy assessment. For each gene family, phylogenies were estimated using 4 methods: (A) distance, (B) parsimony, (C) maximum likelihood, and (D) Bayesian. Shown are the fractions of families with TREEScores greater than 0.80, grouped according to median expectation value (Q1–Q5) and radiation model: recent ($\blacktriangledown$), uniform ($\blacksquare$), and ancient ($\blacktriangle$). Open symbols show families with nonsignificant median expectation values.

families, we estimated phylogenies with *Giardia* open reading frames and their non-*Giardia* homologs using TREE-SCORE (figs. 5D and 6). Despite having a range of sequence similarities that is quite similar to the range of diversity seen in our statistically significant synthetic protein families, *Giardia* phylogenies are significantly less accurate in each of the similarity quartiles (fig. 5D). *Giardia* protein families have TREE-SCORE accuracies similar to those seen in the least statistically significant quintile (Q1) of our synthetic families (fig. 6). For example, 23% of phylogenetic trees built from *Giardia* families were at least 80% correct, compared with the 22% of ancient radiation Q1 families and 46% of Q2 families, using maximum likelihood. Thus, the *Giardia* protein family trees are substantially less accurate than one would expect, based on their statistical significance or expectation value. The lower consistency of the *Giardia* phylogenies presumably reflects the larger number of gaps in these sequences and the shorter ungapped regions (fig. 2D).

In addition to median sequence similarity, we characterized other properties of the *Giardia* protein homologs to see if we could identify a subset of proteins that were more likely to produce accurate evolutionary trees. We characterized the diversity of each family by estimating the statistical significance of sequence similarities of every pairwise alignment in the family and found that all the *Giardia* protein family members shared significant similarity, but families with the stronger sequence similarity produced considerably more accurate trees than the more divergent families (fig.

5B). We did not find a strong correlation between the length of ungapped regions and tree accuracy (data not shown).

Including Truncated Sequences Decreases Phylogenetic Accuracy

Because EST sequences can be readily and cheaply obtained, they are often included in phylogenetic analysis. But EST sequences also contain errors, so we examined the effect of including EST sequences in our phylogenetic analyses. For each of our 2,400 protein families, 1 mRNA sequence was altered using an EST-contig-like error model or an EST-like model (Methods, fig. 7). Phylogenies were estimated using maximum likelihood and evaluated with TREE-SCORE. We sometimes saw a small decrease in the accuracy of phylogenies produced from families with EST-contig-like sequences (fig. 7, dashed line). However, if we limited the truncation to less than 50% of the sequence length, we observed no decrease in phylogenetic accuracy for contig sequences (data not shown). Sequences evolved under a limited truncation EST-like model showed a slight decrease in accuracy for families in Q4 (fig. 7, dotted line) but resembled accurate full-length sequences otherwise. Thus, truncation can reduce accuracy; fortunately, truncated sequences are readily identified in sequence alignments and thus can be excluded.

Strategies for Improving Phylogenetic Accuracy

For many ancient radiation sequence families that share statistically significant similarity, the probability of

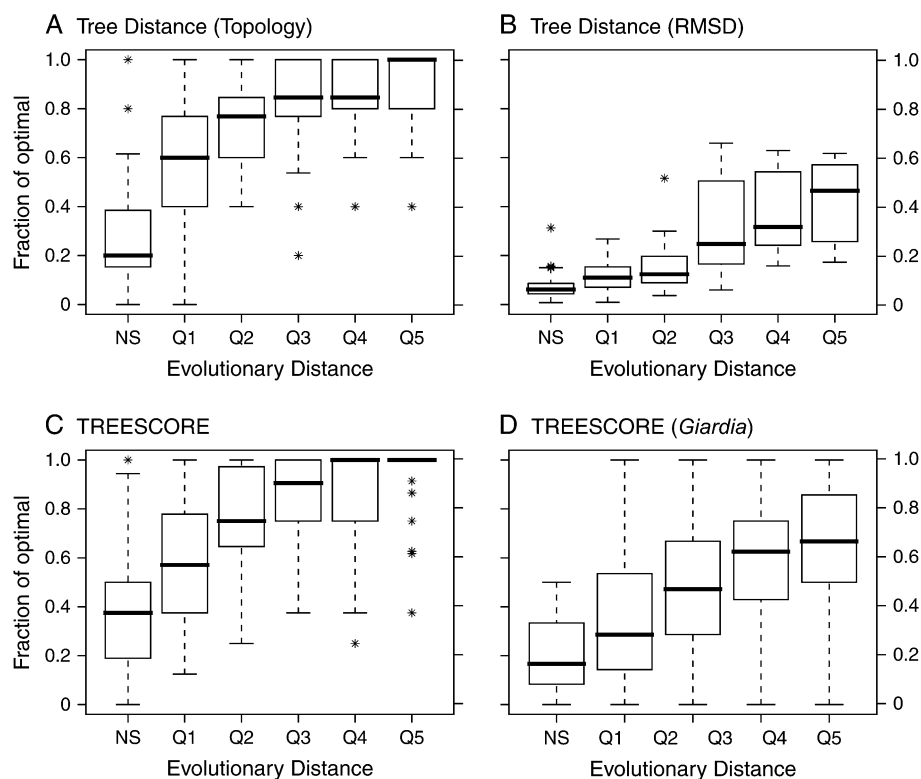


FIG. 5.—Phylogenetic accuracy. Phylogenies were estimated by maximum likelihood, PHYML, and scored according to (A) TREEDIST—symmetric difference, (B) TREEDIST—branch score difference, and (C) TREESCORE. The protein families were divided into 6 groups as in figure 2. (D) TREESCORE distributions for *Giardia* protein, separated into Q1–Q5 quintiles using the expectation value ranges used for the synthetic families in the figures 2–4.

reconstructing an accurate phylogeny was low. Therefore, we examined 3 methods to increase phylogenetic accuracy: using sequence concatenation (fig. 8A and B), correct alignments (fig. 8C), and bootstrapping (fig. 8D).

Increasing sequence size can improve phylogenetic reconstruction (Miyamoto 1985; Kluge 1989; Hillis 1996; Nixon and Carpenter 1996; Graybeal 1998; Hillis 1998; Zwickl and Hillis 2002). Because it is much more difficult to reconstruct phylogenies accurately after ancient radiation, we investigated how the accuracy would improve when sequences were concatenated. We examined 2 different concatenation strategies. First, 5 families that shared the same radiation model and target evolutionary distance, from each of the 48 model trees, were concatenated to create 10 supermatrices per model tree (fig. 8A). Phylogenies were estimated with maximum likelihood and evaluated with TREESCORE. For families evolved using the ancient radiation model, the number of accurate phylogenies increased significantly, from about 63% to 81% correct (fig. 8A, A-I vs. A-C). In the already more accurate phylogenies produced from the uniform and recent radiation families, smaller improvements were seen.

We were concerned that tree accuracy had improved with concatenation because the more distant families were combined with less distant families. Hence, we also concatenated sequences that were evolved under the same topology to the same final evolutionary distances (fig. 8B). Because there were fewer families that met these criteria, only 2 sequences were concatenated. Again, the number

of accurate phylogenies increased significantly. For sequences in quintile Q2, the number of accurate phylogenies increased 50% (fig. 8B). Protein families in the 3 least distant quintiles showed about a 10% increase in accuracy when 2 sequences were concatenated.

Phylogenetic reconstruction begins with multiple sequence alignments, so we asked whether correct multiple sequence alignments produced substantially better phylogenies than calculated alignments. We constructed true alignments based on the positions of insertions and deletions of each sequence and their ancestor sequences. Phylogenies were estimated and scored as in figure 4C. We would expect that for families with poor estimated alignments, NS and Q1, phylogenetic accuracy would increase given the true alignment. Yet, although in some instances we were able to reconstruct a more accurate phylogeny using the true alignment, the differences were minimal even for the most distantly related protein families (Q1, fig. 8C).

In addition, because protein families from our ancient radiation model often produced incorrect trees, we sought a measure of tree reliability that would help us identify the trees that are more likely to be correct. Bootstrap analysis and Bayesian posterior probabilities are used in phylogenetic analysis to determine the robustness of a phylogenetic prediction, so we used these scores to stratify the protein families examined in our analysis. Bootstrap scores were calculated by averaging the bootstrap values from 100 sample data sets of the internal nodes. Posterior probability scores were calculated by averaging the posterior

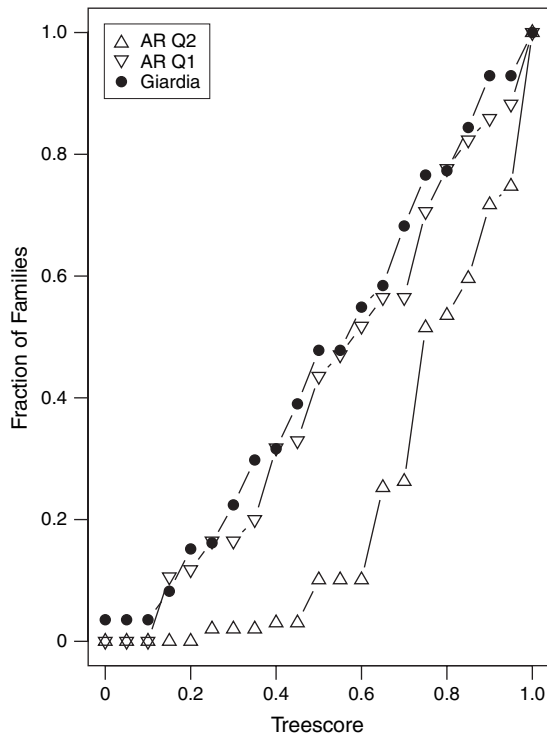


FIG. 6.—Tree accuracy with simulated proteins and *Giardia* protein families. The cumulative fraction of protein families with TREEScores less than or equal to X values. Ancient radiation families in quintile 2 (AR Q2, $\triangle$), ancient radiation families in the least similar quintile (AR Q1, ∇), and *Giardia* protein families ($\bullet$) are shown. Trees were reconstructed using maximum likelihood.

probabilities of the internal nodes. Trees with average bootstrap scores greater than 0.80 across all taxa have more phylogenies that are accurate (fig. 8D, triangle), particularly for families in Q2 and Q3 (70% increase). Similarly, trees with the lowest variance among their bootstrap values at the internal nodes were accurate more often (data not shown). Overall, when families that do not have a bootstrap score greater than 0.80 were removed, accuracy increased from 56% to 62%. Similarly, trees with the highest average posterior probabilities were more likely to be accurate (data not shown).

Discussion

We have explored the limits of phylogenetic reconstruction by assessing the probability of estimating a correct phylogenetic tree for diverse protein families evolved using different radiation models. As expected, as sequence similarity increases, multiple sequence alignment accuracy increases. Phylogenies of protein families evolved under a uniform radiation model are estimated accurately at diverse DNA distances using various phylogenetic algorithms (Hall 2005). However, Hall's study did not examine accuracy under more challenging radiation models. In this study, about 90% of alignments and 80% of phylogenies from families with a uniform or recent radiation evolutionary history that share statistically significant sequence similarity (i.e., those that would have been identified by a Blast search) are accurate. As expected (Fiala and

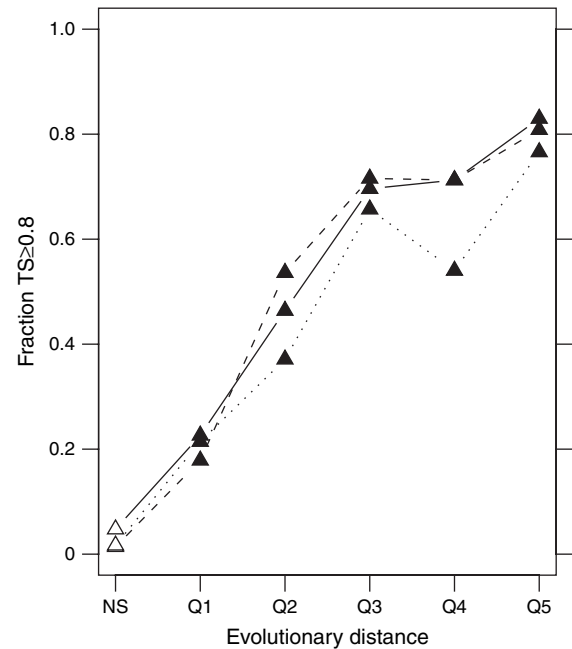


FIG. 7.—Phylogenetic accuracy of simulated EST families. Phylogenies created using maximum likelihood with an EST-contig-like family member (dashed line) and minimally truncated EST-like family member (dotted line) and without either (solid line). Shown is the fraction of families with TREEScore ≥ 0.80 . Filled symbols represent families with a significant median expectation value and therefore identified by sequence similarity searches.

Sokal 1985; Salisbury 1999), although protein families from uniform and recent radiation tree topologies frequently produced accurate estimates of the original phylogeny, phylogenies of proteins from ancient radiations were more difficult to reconstruct accurately. We were surprised to find that 20–30% of the phylogenies from closely related protein families with short ancient branch lengths (ancient radiation) were estimated inaccurately, despite uniformly high pairwise sequence similarity within the family (fig. 4C, upright triangle, Q4, Q5). This result suggests that sequences found by a Blast search, with very good expectation values, can frequently fail to produce correct phylogenies simply because of ancient radiation. Thus, LGT need not be invoked to explain inaccurate or inconsistent phylogenies. Short ancient branches, which can produce LBA, combined with distant protein homologs, and a high frequency of gaps, may be the simplest explanation for the inconsistent phylogenies produced by *Giardia* protein families.

Including ESTs in phylogenetic tree reconstructions reduced accuracy only slightly when the EST sequence covered more than half the length of the protein family. Likewise, when we estimated trees using true multiple sequence alignments, phylogenetic accuracy improved only slightly. Our results are consistent with work by Hall (2005), which shows that topological accuracy does not increase when correct alignments are used. However, branch-length accuracy is increased with accurate alignments (Hall 2005). Although accurate alignments are important for creating accurate trees, other factors, such as sequence similarity and short ancient branch lengths, have much more impact on phylogenetic accuracy. In challenging cases such

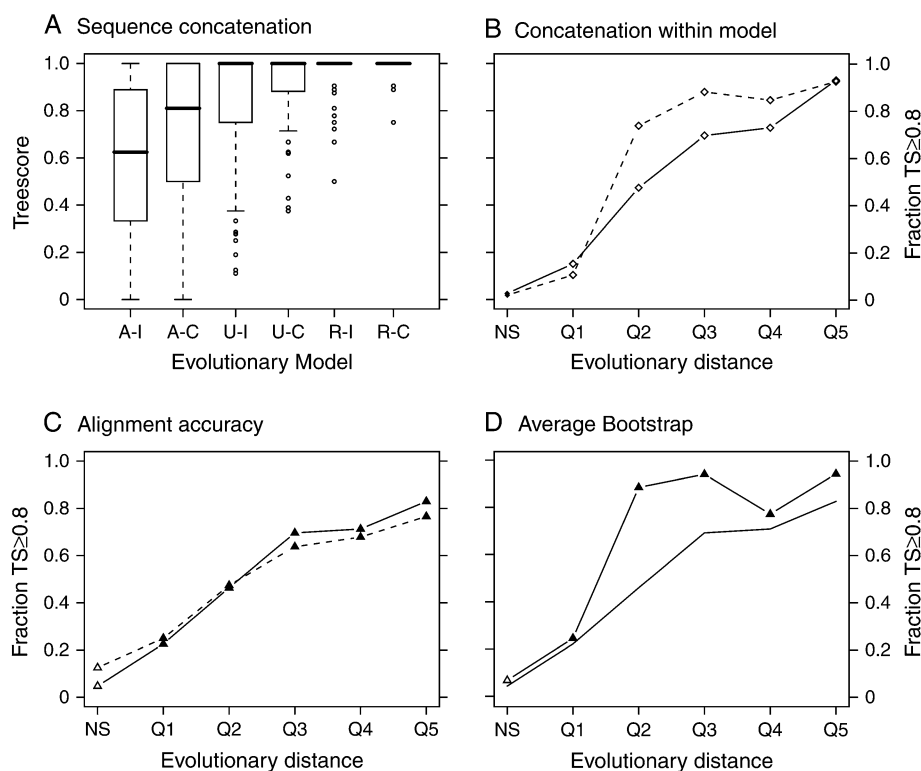


FIG. 8.—Strategies for improving phylogenetic accuracy. The effect of gene concatenation (A, B) and using true alignments (C) on phylogenetic accuracy was examined. Bootstrap analysis was also performed (D). Ancient radiation gene families were grouped as in figure 2B. Filled symbols represent families with a significant median expectation value; members of these families would be identified by sequence similarity searches. Trees were reconstructed using maximum likelihood. (A) Box plots summarize the distribution of TREEScores. A-I (ancient radiation—individual gene), A-C (ancient radiation—concatenated genes), U-I (uniform divergence—individual), U-C (uniform divergence—concatenated), R-I (recent radiation—individual), and R-C (recent radiation—concatenated). As in figure 3, the boundaries of the box indicate the first and third quartile, with the median shown inside the box. The lines outside the box extend 1.5 times the interquartile distance, and circles indicate outliers beyond that range. (B) Gene families were concatenated from families of the same topology and evolutionary distance group. Shown is the fraction of families with accurate phylogenies, classified by a TREEScore greater than 0.80, from individual genes (solid line) and concatenated genes (dashed line). (C) Ancient radiation phylogenies reconstructed from the true multiple sequence alignment (dashed line) are compared with reconstructions from T-COFFEE alignments (solid line). (D) Tree robustness was measured from average bootstrap values for 100 sample trees of internal nodes (bootstrap score). Families with bootstrap scores greater than median ($\triangle$ and $\blacktriangle$) compared with all trees (solid line).

as families with long recent branches, sequence concatenation and choosing trees with high robustness scores (bootstrap values and posterior probability) can improve the probability of accurately estimating the trees.

Because simulation studies are performed under well-specified models, our results almost certainly overestimate our ability to reconstruct accurate phylogenies from biological data. Thus, our estimates of reconstruction accuracy must be seen as best case values; more complex models of rate variation across sequences, and a more comprehensive sample of tree topologies and radiation histories, might allow a more reliable estimate phylogeny accuracy for a given level of sequence similarity. For similar levels of sequence similarity, the *Giardia* families produce much less accurate trees than our simulated families (fig. 5D). The large differences in ungapped alignment coverage between our synthetic families and *Giardia* families suggest that a more realistic model for protein families would contain a domain insertion/deletion model, perhaps with domains that evolve at different rates from the surrounding sequence. Because *Giardia* protein families do show a range of tree accuracies similar to our Q1 synthetic families, it may be possible to develop a more accurate synthetic evolution

model that reproduces the multiple alignment ungapped coverage and has a range of TREEScore accuracies similar to *Giardia*. This model might allow a more accurate prediction of phylogeny reconstruction from sequence similarity and ungapped alignment coverage.

Of course, some of the inaccurate *Giardia* phylogenies may reflect biological phenomena (i.e., LGT or recombination) and possible human errors, such as including paralogous genes or proteins that do not share similar domain composition. But the observation that synthetic protein families with a simple phylogenetic history can produce a similar range of reconstruction accuracies (fig. 6) suggest that a computational, rather than biological, explanation is sufficient to explain the inconsistencies.

Species trees based on molecular data are rarely estimated with individual proteins. Now that many prokaryotic and model eukaryotic genomes have been completed and several groups are producing high-throughput EST sequences, phylogenetic analysis should be performed not on individual protein families but by concatenating several sequences. To improve accuracy further, phylogenies should be estimated with organisms that have close relatives in the protein family and all sequences should share

statistically significant sequence similarity ($E(\cdot) \leq 10^{-3}$) with all other members of the family. And, in studies of phylogenies with many protein families, families with higher bootstrap values across the tree are more likely to produce accurate phylogenies.

Acknowledgments

This work was supported by National Institutes of Health grant AI1058054 to M. Sogin. We thank S. Huse for help with the analysis of the *Giardia* phylogenetic tree data and M. Sogin for his comments on the manuscript. *C. reinhardtii*, *P. sojiae*, and *P. ramorum* sequence data were produced by the US Department of Energy Joint Genome Institute, <http://www.jgi.doe.gov/> and are provided for use in this publication/correspondence only.

Funding to pay the Open Access publication charges for this article was provided by the NIH grant AI1058054.

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–10.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–402.
- Benner SA, Cohen MA, Gonnet GH. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol* 229:1065–82.
- Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol* 16:817–25.
- Campbell AM. 2000. Lateral gene transfer in prokaryotes. *Theor Popul Biol* 57:71–7.
- Dayhoff MO, Eck RV, Park CM. 1972. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*. Volume 5. Silver Spring, MD: National Biomedical Research Foundation. p 89–99.
- Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*. Volume 5. Silver Spring, MD: National Biomedical Research Foundation. p 345–52.
- de la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* 8:128–33.
- Doolittle WF. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14:307–11.
- Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–9.
- Eck RV, Dayhoff MO. 1966. *Atlas of protein sequence and structure*. Silver Springs, MD: National Biomedical Research Foundation.
- Eisen JA. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* 10:606–11.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Biol* 27:401–10.
- Felsenstein J. 1985. Confidence-limits on phylogenies—an approach using the bootstrap. *Evolution* 34:783–91.
- Felsenstein J. 1993. *Phylogeny Inference Package (PHYLP)*. Version 3.5. Seattle, WA: University of Washington.
- Fiala KI, Sokal RR. 1985. Factors determining the accuracy of cladogram estimation—evaluation using computer-simulation. *Evolution* 39:609–22.
- Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155:279–84.
- Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 47:9–17.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
- Hall BG. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol Biol Evol* 22:792–802.
- Hillis DM. 1996. Inferring complex phylogenies. *Nature* 383:130–1.
- Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol* 47:3–8.
- Hirt RP, Logsdon JM Jr, Healy B, Dorey MW, Doolittle WF, Embley TM. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci USA* 96:580–5.
- Jain R, Rivera MC, Moore JE, Lake JA. 2003. Horizontal gene transfer accelerates genome innovation and evolution. *Mol Biol Evol* 20:1598–602.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–82.
- Kluge AJ. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst Zool* 38:7–25.
- Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11:459–68.
- Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc Natl Acad Sci USA* 93:1930–4.
- Lopez-Garcia P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci* 24:88–93.
- Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2.
- Miyamoto MM. 1985. Consensus cladograms and general classifications. *Cladistics* 1:186–9.
- Morrison HG, McArthur AG, Adam RD, Aley SB, Gillin FD, Olsen GJ, Sogin ML. 2004. The *Giardia lamblia* Genome Database. Available from: www.mbl.edu/Giardia.
- Nixon KC, Carpenter JM. 1996. On simultaneous analysis. *Cladistics* 12:221–41.
- Notredame C, Higgins DG, Heringa J. 2000. T-COFFEE: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205–17.
- Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
- Olsen GJ. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp Quant Biol* 52:825–37.
- Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444–8.
- Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* 46:24–36.
- Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci* 53:131–47.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–4.
- Salisbury BA. 1999. Misinformative characters and phylogeny shape. *Syst Biol* 48:153–69.

- Simossis VA, Kleinjung J, Heringa J. 2005. Homology-extended sequence alignment. *Nucleic Acids Res* 33:816–24.
- Van de Peer Y, Ben Ali A, Meyer A. 2000. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. *Gene* 246:1–8.
- Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11:367–72.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–6.
- Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* 15:1600–11.
- Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 51:588–98.

William Martin, Associate Editor

Accepted July 31, 2006