Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences

Barry G. Hall

Biology Department, University of Rochester

A biologically realistic method was used to simulate evolutionary trees. The method uses a real DNA coding sequence as the starting point, simulates mutation according to the mutational spectrum of *Escherichia coli*—including base substitutions, insertions, and deletions—and separates the processes of mutation and selection. Trees of 8, 16, 32, and 64 taxa were simulated with average branch lengths of 50, 100, 150, 200, and 250 changes per branch. The resulting sequences were aligned with ClustalX, and trees were estimated by Neighbor Joining, Parsimony, Maximum Likelihood, and Bayesian methods from both DNA sequences and the corresponding protein sequences. The estimated trees were compared with the true trees, and both topological and branch length accuracies were scored. Over the variety of conditions tested, Bayesian trees estimated from DNA sequences that had been aligned according to the alignment of the corresponding protein sequences were the most accurate, followed by Maximum Likelihood trees estimated from DNA sequences and Parsimony trees estimated from protein sequences.

Introduction

Phylogenetic analyses are now routinely used by nonsystematists as tools for understanding a variety of biological processes. Applications of phylogenetic methods include understanding genome organization, epidemiology, predicting protein functions, and deciding which genes to analyze in comparative studies. Typically, such users are less interested in understanding the historical relationships among species than they are in understanding the relationships among the sequences of the genes and proteins that they subject to phylogenetic analysis. As they become aware of the variety of phylogenetic methods that are available, they are faced with the difficult problem of choosing the most appropriate method for their purposes. Users are interested in the relative accuracies of the methods, as well as in the trade-offs among accuracy, computational speed, and ease of use of the programs that implement the various methods. Although there is a considerable literature on the comparison of various methods, there is little in that literature that provides clear guidance to the casual user of phylogenetic analysis.

Comparison of the various methods requires the generation of a data set from a known or true tree, applying the various phylogenetic methods to that data set, and then comparing the estimated trees with the true tree to determine the accuracy with which each method estimates the tree.

The most reliable true trees come from experimental evolution systems in which a population of organisms, typically bacteriophage, is periodically divided into lineages (Hillis et al. 1992; Hillis, Huelsenbeck, and Cunningham 1994; Bull et al. 1997). The shape of the phylogeny, i.e., the order of branching events and the time between branches, is determined by the experimenter, but the evolutionary changes depend upon the properties of the experimental system itself. The number and nature of the changes is determined from the sequences of the gene(s) of interest at each branch point. Such experimental studies are resource intensive, and most are limited to a small number of lineages.

Key words: simulation, alignment, method comparison, indels, branch lengths, topology.

E-mail: drbh@mail.Rochester.edu.

Mol. Biol. Evol. 22(3):792–802. 2005
doi:10.1093/molbev/msi066
Advance Access publication December 8, 2004

The alternative is computer simulations of sequence evolution to generate true trees. Typically a random sequence is evolved along a model tree by random substitutions according to some evolutionary model (Li 1997), such as the Kimura Two Parameter model (Kimura 1980). In some cases the model tree involves a constant rate of evolution, and in others the rate is variable, but even in the case of variable rates the variation is typically according to a simple pattern (Saitou and Imanishi 1989). However, in at least one study (Kuhner and Felsenstein 1994) the model trees were randomly constructed, resulting in both a variety of different topologies and variable evolutionary rates so that the true trees were comparable to those seen in real data.

Those simulations suffer a common set of problems. They typically consider only a small number (4–32) of taxa. They may be biased toward one method or another by the evolutionary model that is chosen. Whatever evolutionary model is used, it certainly oversimplifies the reality of the substitution process. Indeed, modeling evolution as a process of substitution confounds two distinct processes, mutation and selection, whose outcome is the real substitution pattern. Mutation is a complex process that ultimately generates a ''spontaneous mutational spectrum,'' the relative proportions of the various kinds of mutations (base substitutions, insertions, and deletions) that occur. The basis of the base substitution process is fairly well understood: it is the result of DNA polymerase incorporation errors combined with the failures of the various repair systems to eliminate all of those errors (Schaaper and Dunn 1991). The basis of insertion and deletion mutations is much less well understood. About half of deletions can be attributed to local repeat motifs, but the basis of the remaining half is not well understood. Once a mutation has occurred selection and drift combine to determine which mutations are fixed into populations. Because most mutations that result in amino acid substitutions are deleterious, and because of the redundant nature of the genetic code, the mutation spectrum can be very different from the substitution spectrum.

Probably the most serious common failing of most computer simulated trees is their failure to incorporate deletions and insertions during their evolution. The resulting terminal-taxon sequences require no alignment prior to using them as the data to which various phylogenetic

methods will be applied for comparison. In reality, sequences must be aligned, and the qualities of the resulting trees depend strongly on the qualities of the alignments. The failure to require the alignment step considerably reduces the confidence we can have in using the resulting comparisons of methods as a guide to making decisions about which methods to use with real DNA. I am aware of only one program, ROSE (Stoye, Evers, and Meyer 1998), that incorporates insertions and deletions during the simulated evolution of sequences.

Once the data set is created by simulation, the phylogenetic methods to be compared are applied to estimate trees, and the estimated trees are compared with the true tree to produce some measure of the accuracy of the method. Most studies consider accuracy only in terms of the topologies of the estimated trees, and the most common measure of accuracy is the percent of the time among replicates that the topology of the estimated tree is identical to the topology of the true tree. That is not a very useful measure because it yields no indication of how close the typical estimated tree is to the true tree. If the estimated tree differs from the true tree by a single branch or by many branches it is scored identically as non-identical. In fact, we are less interested in how often an estimated tree is perfect among 1000 replicates than we are in how good is the estimated tree likely to be.

There is a notable exception to that measure of accuracy. Kuhner and Felsenstein (Kuhner and Felsenstein 1994) measured quality of the estimated trees, not just the quantity that were perfect. Theirs was also the only study I am aware of that included a measure of the accuracy of branch lengths. While most systematists are primarily interested in accuracy of topologies, others are equally interested in the accuracies of branch lengths. Finally, theirs is also the only study that I am aware of that considered the relative speeds of the various methods.

The study reported here has several purposes: first, to develop a biologically realistic method of sequence evolution simulation, a method that separates mutation from selection and that incorporates insertions and deletions; second, to develop a method of comparing estimated trees with a true tree that scores the quality of the topologies estimated for the trees rather than the quantity of trees that are perfect, and that scores the quality of the estimated branch lengths; third, to apply those methods to comparing the accuracies of trees based on protein sequences with those based on DNA sequences when using Neighbor Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML), and the Bayesian method to estimate those trees.

Methods

Simulation of DNA Sequence Evolution

Sequence evolution was performed by the program EvolveAGene 2.2 (Hall 2004*e*), written in C and executed on a Macintosh G4 computer. Details of the program are discussed in the *Results*. The simulations were all initiated by the coding region of *XisC*, a 1494 base pair hupL element site-specific recombinase from an Anabena species, GenBank Accession number U08014, from which the termination codon had been removed.

Sequence Alignments

Protein sequences were aligned using ClustalX 1.83 (Thompson et al. 1997) with pairwise gap penalties of 35 for gap opening and 0.75 for gap extension, and multiple alignment penalties of 15 for gap opening and 0.3 for gap extension.

DNA sequences were either aligned ''directly'' with ClustalX using the default gap opening penalties of 15 and gap extension penalties of 6.66 for both the pairwise and multiple alignment stages, or they were aligned, as indicated, according to the corresponding protein sequence alignment using CodonAlign 2.0 (Hall 2004*a*) as previously described (Hall 2004*b*). In neither case were alignments optimized either by modifying global gap penalties or by modifying local gap penalties for selected ranges of residues.

ClustalX calculates a quality score (Q-score) for each site in the alignment and displays those scores as a histogram below the alignment pane. The Q-scores for alignments were saved to text files and the program TuneClustalX 1.01 (Hall 2004*c*) was used to calculate the average Q-score as a measure of the overall quality of the alignments.

Estimation of Phylogenetic Trees

Neighbor Joining, Parsimony, and Maximum Likelihood trees were estimated using PAUP*4.0b10 (Swofford 2000) Bayesian trees were estimated using MrBayes 3.0b4 (Huelsenbeck and Ronquist 2001).

Tree Comparisons

Estimated trees were compared with their corresponding true trees by the program CompareTrees 1.01, as described in detail in the *Results*. CompareTrees was written in C and executed on a Macintosh G4 computer. Details of the program are also discussed in the *Results*.

Computer Programs

The programs CodonAlign 2.0, TuneClustalX, EvolveAGene, and CompareTrees are available from the author on request. CodonAlign 2.0 is available for Macintosh and Windows platforms, and the source code is available to be compiled for Unix machines. The other programs are available only for the Macintosh platform.

Results

The Simulation Process

The EvolveAGene program was designed to generate phylogenetic trees by simulating the evolution of sequences in a biologically realistic fashion. The program takes as its input a real DNA coding sequence. That sequence forms the root node of the tree. The user chooses the number of taxa (external nodes) on the tree, either 2, 4, 8, 16, 32, 64, or 128 taxa. The present study was limited to a maximum of 64 taxa. The user chooses the average length of the branches on the tree, where the length is defined as the number of sequence changes between nodes, i.e., the number of accepted mutations.

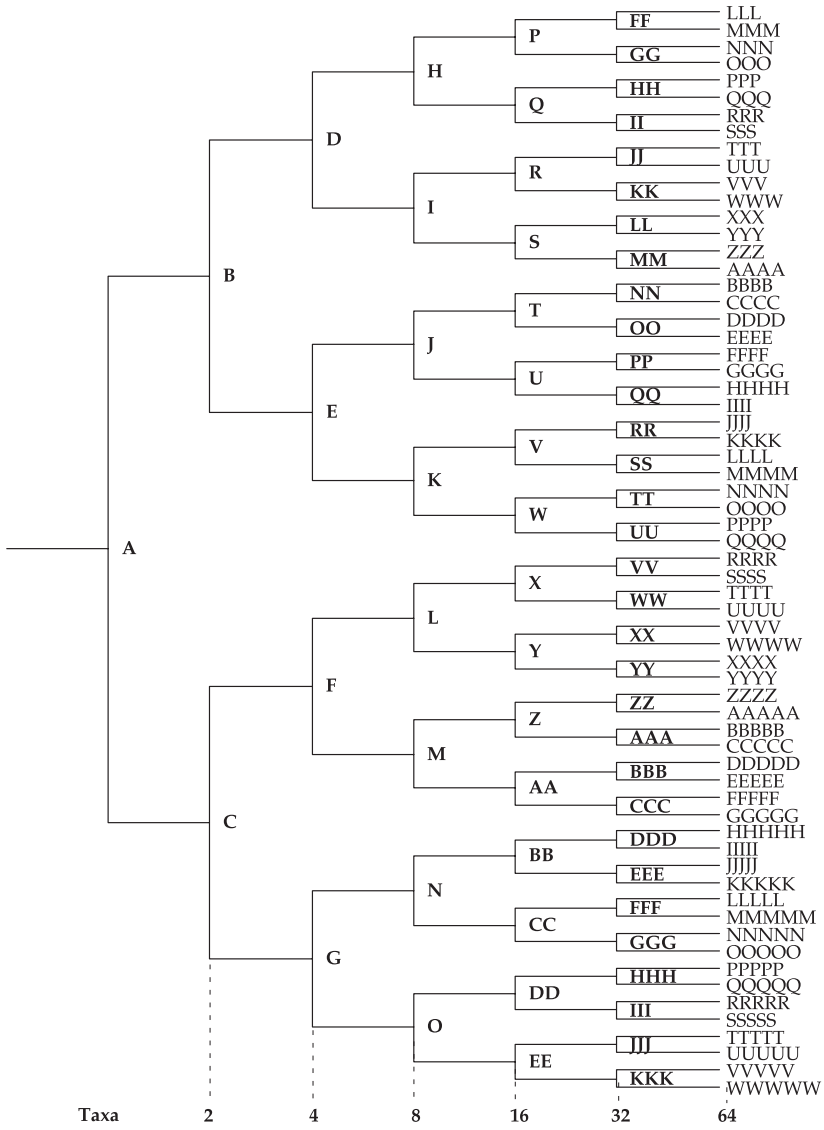

FIG. 1.—Node-naming convention for simulated trees.

The program creates a strictly bifurcating, cladistically symmetric, tree starting from the root. For each branch, the length is a random number between 0 and twice the average branch length defined by the user. This process means that the evolutionary rate varies throughout the tree. Figure 1 shows the naming convention for the nodes, and figure 2 shows a typical tree of 64 taxa with an average of 250 changes per branch. The arrows in figure 2 illustrate how the strictly bifurcating tree can include near-trichotomies that result from near-zero branch lengths.

Once the tree is established, the program evolves the sequences along the tree, starting from the input root sequence. The user determines the probability that an amino acid replacement mutation will be accepted, and the probabilities that insertion and deletion mutations will be accepted.

At each attempt to change the sequence, a random site in the DNA sequence is chosen and a mutation of that site is attempted according to the mutational spectrum of *Escherichia coli*.

A mutational spectrum shows the proportions of the various kinds of spontaneous mutations in a gene—i.e., base substitutions; insertions; and deletions; among base substitutions, the proportions of the six kinds of possible base changes; and among insertions and deletions, the proportions of indels of different lengths. Note that these are not the proportions of substitutions that are found among existing genes in the populations; they are the proportions of spontaneous mutations that are experimentally determined to occur in target genes. The most reliable studies of mutational spectra involve determining loss-of-function mutations in genes where loss of function is easily selected. The two most thoroughly studied target genes are *lacI* (Glickman, Burns, and Fix 1986) and *ebgR* (Hall 1999). Taken together, those studies show that 61% of spontaneous mutations are base substitutions, 33% are deletions, and 6% are insertions. Eight percent of base substitutions are AT to GC, 40% are GC to AT, 29% are GC to TA, 4% are GC to CG, 10% are AT to CG, and 10% are AT to

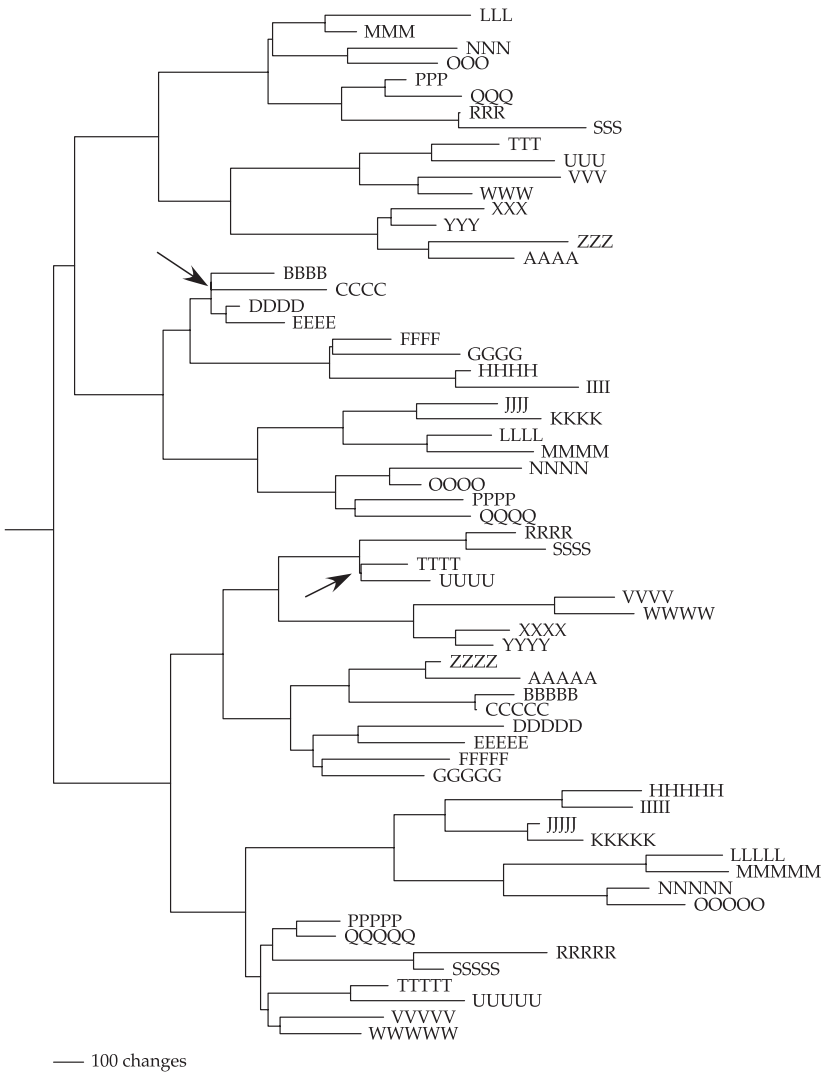

FIG. 2.—Phylogram of a typical true tree of 64 taxa with an average of 250 changes (accepted mutations) per branch.

TA; thus, transitions occur almost as frequently (48%) as transversions (52%). This is very different from typical substitution spectra where transitions outnumber transversions by about 1.5:1 (Li 1997). The length distributions of insertions and deletions do not form a simple pattern, in part because the number of indels observed in those studies was small. Although we know a good deal about the molecular basis of the base substitution spectra, we know relatively little about the molecular basis of indels beyond the fact that about 50% of indels are associated with repeat sequences (Schaaper, Danforth, and Glickman 1986). About 50.9% of deletions involve a single base, with all other lengths observed being 1%-3% each. For the purpose of the simulation, I have taken all deletion lengths >1 base as having equal probabilities of 2.3%, with deletions limited to a maximum of 23 bases. Similarly, insertions of 1 base have a probability of 52.6%; two-base insertions, a probability of 9.3%; and lengths >2, probabilities of 4.8% up to a maximum of 11 bases.

When a random site is chosen by EvolveAGene it is mutated according to the *E. coli* spectrum described above. If the mutation is an indel whose length is not an integer multiple of three, the mutation is not accepted on the grounds that such mutations result in frameshifts that are almost always strongly selected against because they result in complete loss of protein function. If the indel is an integer multiple of three, it is accepted or rejected according to the probability specified by the user. When a mutation is not accepted, the sequence remains unchanged, another random site is chosen, and another attempt to introduce a mutation is made. If the mutation is a base substitution and the encoded amino acid is unchanged—i.e., the mutation is silent—the mutation is accepted and incorporated into the sequence. If the mutation results in a nonsense (chain termination) codon, the mutation is rejected. If the amino acid is changed to a different amino acid—i.e., the mutation was nonsynonymous—the mutation is accepted or rejected according to the probability specified by the user. As a result, the probability of accepting an amino acid substitution is equivalent to the dN/dS ratio across the tree. That equivalence has been confirmed for several trees by Yang's Codeml program of the PAML suite (Yang 1997) (results not shown).

When the number of accepted mutations is that specified by the length of the current branch, the sequence is saved

Table 1
Tree Scores [a]

Average Branch Length	Mean Protein Alignment Q-Score	NJ Protein	Parsimony Protein	Bayesian Protein	NJ DNA-DD	Parsimony DNA-DD	ML DNA-DD	Bayesian DNA-DD	NJ DNA-CA	Parsimony DNA-CA	ML DNA-CA	Bayesian DNA-CA	Mean DNA Alignment Q Score
						Eight taxa							
50	74.81	0.893	0.881	0.837	0.782	0.822	0.936	0.795	0.793	0.836	0.934	0.789	83.29
100	63.20	0.817	0.904	0.904	0.677	0.689	0.876	0.877	0.688	0.702	0.916	0.914	75.16
150	52.71	0.746	0.808	0.847	0.516	0.606	0.827	0.828	0.534	0.622	0.914	0.776	68.98
200	41.98	0.806	0.817	0.723	0.567	0.588	0.786	0.789	0.573	0.590	0.896	0.778	64.64
250	38.10	0.662	0.726	0.836	0.459	0.504	0.684	0.688	0.454	0.493	0.820	0.663	59.72
						Sixteen taxa							
50	66.45	0.867	0.912	0.879	0.728	0.740	0.848	0.842	0.744	0.751	0.839	0.831	78.05
100	44.79	0.782	0.827	0.777	0.539	0.647	0.784	0.783	0.541	0.658	0.915	0.916	62.61
150	43.85	0.823	0.809	0.856	0.605	0.649	0.824	0.825	0.617	0.647	0.881	0.881	62.01
200	32.86	0.684	0.782	0.802	0.473	0.518	0.743	0.743	0.474	0.533	0.746	0.823	55.96
250	20.10	0.641	0.750	0.556	0.418	0.504	0.736	0.739	0.475	0.505	0.718	0.714	47.33
						Thirty-two taxa							
50	57.22	0.827	0.879	0.823	0.715	0.793	0.853	0.846	0.745	0.791	0.913	0.903	71.04
100	45.78	0.779	0.843	0.843	0.563	0.665	0.793	0.788	0.575	0.670	0.873	0.878	63.61
150	29.22	0.767	0.823	0.732	0.535	0.626	0.762	0.760	0.591	0.618	0.814	0.809	52.99
200	18.91	0.640	0.753	0.655	0.473	0.522	0.805	0.805	0.468	0.557	0.793	0.761	46.18
250	14.09	0.529	0.702	0.409	0.380	0.457	0.742	0.745	0.372	0.430	0.677	0.670	42.08
						Sixty-four taxa							
50	50.82	0.857	0.868	0.803	0.697	0.822	0.841	0.845	0.734	0.814	0.852	0.855	66.37
100	33.20	0.153	0.835	0.728	0.585	0.711	0.813	0.821	0.587	0.703	0.845	0.852	55.44
150	18.22	0.722	0.761	0.597	0.483	0.578	0.649	0.660	0.540	0.582	0.792	0.800	44.22
200	12.40	0.674	0.758	0.521	0.494	0.549	0.739	0.764	0.483	0.557	0.705	0.721	38.76
250	11.416	0.5718	0.6835	0.4592	0.4238	0.4785	0.6694	0.697	0.438	0.502	0.714	0.736	37.97

[a] The input (root) sequence was *XisC*, 1494 bp. Probabilities of accepting non-synonymous substitutions, insertions and deletions were all 0.1.

according to the naming convention in figure 1. Internal node sequences are saved to a separate file from terminal node (taxon) sequences. Both internal node and taxon sequences are translated and saved to separate protein sequence files. Files containing the true tree with branch lengths in numbers of DNA sequence changes and with branch lengths in the number of protein sequence changes are saved. Multiple changes at the same site are each counted in those branch lengths.

The result of the simulation is a biologically realistic tree that is initiated by a real coding sequence, a tree in which mutation and selection are separate processes, a tree that is based on the spontaneous mutation spectrum of a real organism, and a tree that includes both insertion and deletion mutations.

The Alignment Process

Before being used to estimate phylogenetic trees, real sequences must be aligned, and it is a truism that the quality of a tree is no better than the quality of the alignment used to estimate that tree (Hall 2004*b*).

Taxon protein sequences were aligned with ClustalX 1.83 as described in *Methods*, and the average Quality Scores were calculated.

Two methods were used to align the DNA sequences. DNA sequences were ''directly'' aligned using ClustalX 1.83 as described in *Methods*. Trees based on those alignments are specified in tables 1–4 as ''DNA-DD.'' DNA sequences were also aligned according to the protein sequence alignments using CodonAlign 2.0 (Hall 2004*b*). Trees based on those alignments are specified as ''DNA-CA.'' The rationale for CodonAligned DNA sequences is that alignment algorithms introduce gaps that maximize the alignment score. Those gaps are intended to represent historical insertions and deletions during the phylogenetic histories of the sequences. Often, gaps that maximize alignment scores result in frameshifts when the gapped sequences are translated; thus the translated aligned sequences bear little resemblance to the real protein sequences. Had such frame shifting gaps actually arisen during the history of a protein, the resulting alleles would almost certainly have been eliminated by purifying selection.

Tree Comparisons

Phylogenetic trees were estimated from protein sequences, from directly aligned DNA sequences (DNA-DD), and from CodonAligned DNA sequences (DNA-CA) derived from each data set. Aside from comparing the various phylogenetic methods, one of the major purposes of these studies was to determine whether trees estimated from DNA sequences are more or less accurate than trees estimated from the corresponding protein sequences. Another purpose was to determine whether trees estimated from CodonAligned DNA sequences are, in fact, more accurate than trees estimated from directly aligned DNA sequences.

The estimated trees were compared with the True Trees with the program CompareTrees (Hall 2004*d*). CompareTrees calculates three scores for each comparison: a topology score, a branch length score, and a tree score. The topology score is the fraction of clades that are present in the True Tree that are also present in the estimated tree. The branch length score is determined by calculating, for

Table 2
Topology Scores[a]

Average Branch Length	Mean Protein Alignment Q-Score	NJ Protein	Parsimony Protein	Bayesian Protein	NJ DNA-DD	Parsimony DNA-DD	ML DNA-DD	Bayesian DNA-DD	NJ DNA-CA	Parsimony DNA-CA	ML DNA-CA	Bayesian DNA-CA	Mean DNA Alignment Q Score
						Eight taxa							
50	74.81	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	83.29
100	63.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	75.16
150	52.71	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	68.98
200	41.98	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	64.64
250	38.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	59.72
						Sixteen taxa							
50	66.45	1.000	1.000	1.000	1.000	0.929	0.929	0.929	1.000	0.929	0.929	0.929	78.05
100	44.79	1.000	0.952	0.929	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	62.61
150	43.85	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	62.01
200	32.86	0.929	1.000	1.000	0.929	0.929	0.929	0.929	0.929	1.000	0.929	1.000	55.959
250	20.10	1.000	1.000	0.929	0.857	0.929	0.929	0.929	1.000	0.929	0.929	0.929	47.330
						Thirty-two taxa							
50	57.22	0.967	0.978	0.967	1.000	0.983	1.000	1.000	1.000	0.967	1.000	1.000	71.04
100	45.78	1.000	0.978	0.967	1.000	0.933	1.000	1.000	1.000	0.967	0.967	1.000	63.61
150	29.22	1.000	1.000	1.000	0.967	0.950	1.000	1.000	1.000	0.933	0.967	0.967	52.99
200	18.91	0.933	0.967	0.967	0.933	0.917	0.967	0.967	0.933	0.967	1.000	0.967	46.18
250	14.09	0.933	0.950	0.967	0.967	1.000	1.000	1.000	0.967	0.933	0.967	0.967	42.08
						Sixty-four taxa							
50	50.82	0.984	0.976	0.984	0.984	1.000	0.984	1.000	1.000	1.000	0.967	0.984	66.37
100	33.20	0.968	0.973	0.968	0.968	0.984	0.968	0.984	0.968	0.984	0.968	0.984	55.44
150	18.22	1.000	0.976	0.984	0.952	0.919	0.936	0.952	1.000	0.979	0.968	0.984	44.22
200	12.40	0.984	0.981	0.968	0.968	0.968	0.968	0.984	0.952	0.968	0.968	0.984	38.76
250	11.42	0.919	0.944	0.919	0.887	0.919	0.871	0.887	0.936	0.952	0.968	0.984	37.97

[a] The input (root) sequence was *XisC*, 1,494 bp. Probabilities of accepting nonsynonymous substitutions, insertions, and deletions were all 0.1.

each branch that is present in both the True Tree and the estimated tree, the absolute value of the difference in branch lengths, and dividing that by the length of the branch in the True Tree. That number is subtracted from 1 to produce the score for that branch, and those scores are averaged to produce the branch length score for the tree. The rationale for using the absolute value of the difference in branch lengths is that it is no better to overestimate than to underestimate the length of a branch, and that averaging signed branch length differences would result in positive differences canceling negative differences, thereby overestimating the average branch length score. The tree score is the product of the topology score and the branch length score; thus it weights topology and branch lengths equally in determining the accuracy of a tree. The tables show both topology scores and tree scores, and readers who would weight those factors differently may use those scores to calculate the branch length scores and differently weighted tree scores.

Tables 1 and 2 show the results for simulated trees of 8, 16, 32, and 64 taxa with an average of 50, 100, 150, 200, and 250 accepted mutations per branch. For each number of taxa, as the average branch lengths increase, the Q-scores of the alignments decrease. Similarly, for each set of branch lengths, as the number of taxa increases, the Quality Scores decrease. Although the Q-scores can be maximized by optimizing the gap penalties, in large part, the Q-scores reflect the diversity of the sequences. It is therefore not surprising to see that, in general, for each phylogenetic method examined, as the Q-scores decrease accuracies of the trees, as represented by the tree scores, also tend to decrease (table 1). The topology scores also decrease slightly with decreasing Q-scores (table 2), but the effect is slight and not significant. Tables 3 and 4 show the results for five replicate data sets for the most extreme set of conditions, 64 taxa and an average of 250 accepted mutations per branch. Tables 3 and 4 also show that under identical simulation conditions there is considerable variation from one run to the next. The standard errors of the scores allow us to judge how seriously to take differences in tables 1 and 2.

Tree scores for directly aligned DNA sequences (DNA-DD) in table 1 were compared with those of Codon Aligned DNA sequences (DNA-CA) by a paired *t*-test. The tree scores for DNA-CA sequences averaged 0.021 higher than the tree scores for DNA-DD sequences. The difference is highly significant ($P < 0.001$). The same comparison of topology scores (table 2) showed that CodonAligned DNA gave slightly higher (0.0073) topology scores with $P = 0.02$. It is reasonable to conclude that it is preferable to align DNA coding sequences according to the alignment of their corresponding protein sequences.

For both CodonAligned DNA and for protein sequences Neighbor-Joining trees typically have the lowest tree score, and ML trees estimated from CodonAligned DNA (DNA-CA) have the highest scores. Just below the tree scores for ML trees for DNA-CA are Bayesian DNA-CA trees and parsimony trees of protein sequences. Each is significantly worse than ML of DNA-CA ($P = 0.05$ and $P = 0.03$, respectively), but they are not significantly different from each other. The superiority of ML trees based on DNA-CA disappears for the 64-taxon set, for which Bayesian DNA-CA trees are significantly better than ML DNA-CA trees, but not significantly different from parsimony protein trees. The major conclusion

Table 3
Reproducibility of Tree Scores[a]

Replicate	Mean Protein Alignment Q Score	NJ Protein	Parsimony Protein	Bayesian Protein	NJ DNA-DD	Parsimony DNA-DD	ML DNA-DD	Bayesian DNA-DD	Mean DNA Alignment Q Score	NJ DNA-CA	Parsimony DNA-CA	ML DNA-CA	Bayesian DNA-CA	NJ K2P[b] DNA-CA	Weighted Parsimony Tv = 2 DNA-CA
1	12.9	0.590	0.714	0.585	0.438	0.493	0.716	0.728	40.5	0.419	0.493	0.735	0.764	0.634	0.708
2	12.2	0.576	0.633	0.428	0.413	0.447	0.677	0.694	39.6	0.402	0.451	0.675	0.685	0.586	0.601
3	11.4	0.572	0.684	0.459	0.424	0.479	0.669	0.697	38.0	0.438	0.502	0.714	0.736	0.640	0.683
4	11.4	0.594	0.656	0.444	0.407	0.428	0.597	0.620	38.2	0.424	0.472	0.716	0.739	0.659	0.636
5	13.9	0.600	0.642	0.420	0.455	0.477	0.693	0.718	39.4	0.433	0.487	0.669	0.678	0.630	0.642
Mean ± S.E.	12.36 ± 0.48	0.586 ± 0.005	0.666 ± 0.015	0.467 ± 0.030	0.427 ± 0.009	0.465 ± 0.012	0.670 ± 0.020	0.691 ± 0.019	39.1 ± 1.3	0.423 ± 0.006	0.481 ± 0.009	0.702 ± 0.035	0.720 ± 0.017	0.630 ± 0.012	0.654 ± 0.019

[a] Sixty-four taxa, branch lengths were an average of 250 accepted mutations. Probabilities of accepting nonsynonymous substitutions, insertions, and deletions were all 0.1.
[b] K2P: Kimura two-parameter model of nucleotide substitution.

Table 4
Reproducibility of Topology Scores[a]

Replicate	Mean Protein Alignment Q Score	NJ Protein	Parsimony Protein	Bayesian Protein	NJ DNA-DD	Parsimony DNA-DD	ML DNA-DD	Bayesian DNA-DD	Mean DNA Alignment Q Score	NJ DNA-CA	Parsimony DNA-CA	ML DNA-CA	Bayesian DNA-CA	NJ K2P[b] DNA-CA	Weighted Parsimony Tv=2 DNA-CA
1	12.90	0.968	0.976	1.000	0.967	0.952	0.936	0.936	40.5	0.952	0.952	0.984	1.000	0.951	0.952
2	12.18	0.887	0.858	0.887	0.903	0.903	0.903	0.919	39.6	0.887	0.887	0.903	0.919	0.887	0.855
3	11.42	0.919	0.944	0.919	0.887	0.919	0.871	0.887	38.0	0.936	0.952	0.968	0.984	0.936	0.952
4	11.39	0.936	0.911	0.968	0.855	0.839	0.823	0.839	38.2	0.936	0.936	0.952	0.968	0.952	0.936
5	13.92	0.903	0.914	0.952	0.936	0.909	0.936	0.952	39.4	0.903	0.919	0.936	0.936	0.936	0.919
Mean ± S.E.	12.36 ± 0.48	0.923 ± 0.014	0.920 ± .020	0.945 ± .019	0.909 ± .019	0.904 ± .018	0.894 ± .021	0.906 ± .020	39.1 ± 0.5	0.923 ± .012	0.929 ± .012	0.948 ± .014	0.961 ± .015	0.932 ± 0.012	0.923 ± 0.018

[a] Sixty-four taxa, branch lengths were an average of 250 accepted mutations. Probabilities of accepting nonsynonymous substitutions, insertions, and deletions were all 0.1.
[b] K2P: Kimura two-parameter model of nucleotide substitution.

from the data in tables 1 and 2 is that the most reliable method, that which yields the most accurate trees overall, is the maximum likelihood method using CodonAligned DNA.

The tree-estimation methods used for tables 1 and 2 were all implemented with the default settings of the respective methods because those are the settings that are most likely to be employed by a casual user who wants to construct phylogenetic trees from molecular data. Having established that the most demanding conditions tested were 64 taxa with average branch lengths of 250 accepted mutations, and that directly aligned DNA sequences consistently produce less accurate trees than do CodonAligned DNA or protein sequences, it appeared reasonable to consider some variations of the basic methods. Because CodonAligned DNA sequences are clearly preferable to directly aligned DNA sequences, only protein sequences and Codon Aligned DNA sequences were considered. Neighbor Joining for DNA-CA sequences was extended to include the Kimura two-parameter model of nucleotide substitution (Kimura 1980), and Parsimony for DNA-CA sequences was extended to include Weighted Parsimony (Maddison and Maddison 1992) with the transversion weight penalties set at 2. The additional methods were applied to the five replicate data sets of 64 taxa and average branch lengths of 250 accepted mutations (tables 3 and 4, last two columns). Paired *t*-tests showed that, in terms of tree scores (table 3), NJ with the Kimura two-parameter model of nucleotide substitution is significantly more accurate ($P < 0.0001$) than NJ with uncorrected distance, and that weighted parsimony is significantly more accurate than unweighted parsimony ($P < 0.0001$). This is consistent with earlier findings that Weighted Parsimony is more accurate than unweighted parsimony (Hillis, Huelsenbeck, and Cunningham 1994). The topology scores (table 4), on the other hand, were not significantly different from each other.

Those conditions push the limits for producing reliable alignments. Q-scores of the protein alignments averaged only 12.36, and pairwise alignments of the most divergent taxa, those whose common ancestor is the root, showed that the percent identical amino acids (%ID) were in the range of 20% to 30% identity. In that so-called twilight zone of evolutionary relatedness (Doolittle 1981), ClustalX aligns, on average, 80% of residues correctly (Thompson, Plewniak, and Poch 1999). Below 10% ID, it aligns <50% of residues correctly (Thompson, Plewniak, and Poch 1999).

In terms of topology, for the data in table 4, all of the methods performed well with topology scores >0.92, but the Bayesian DNA-CA method performed better than all other methods ($P < 0.05$ in all paired *t*-tests).

In terms of tree scores, for the data in table 3, the Bayesian DNA-CA method performed better than all other methods ($P < 0.01$ in all paired *t*-tests).

As pointed out above, the conditions under which evolution of these sequences have been simulated result in sequence sets that push the limits of the ClustalX alignment algorithm. Those conditions should become less stringent if fewer deletion and insertion mutations are accepted. Accordingly, a set of five replicate data sets was created, in which the probability of accepting an insertion or deletion was reduced from 0.1 to 0.025. Thus, the conditions were identical to those in tables 3 and 4 except for the lower probability of accepting an indel mutation. Table 5 shows that those conditions increase the average Q-score almost twofold. Under those less stringent conditions, all of the topology scores were improved, and all of the tree scores were improved (compare table 5 with tables 3 and 4). In terms of the tree scores, the most dramatic differences between the less stringent and the more stringent conditions were that, under the less stringent conditions, Bayesian protein trees and Bayesian DNA-CA trees were significantly more accurate than trees constructed by any other method, but they were not significantly different from each other. The accuracy of Bayesian protein trees appears to be particularly sensitive to the quality of the alignments. In terms of the topology scores, the five most accurate methods were not significantly different from one another.

Discussion

A biologically realistic method was used to simulate evolutionary trees. In all cases the root (input) sequence was a real DNA coding sequence. The mutation process was simulated based on the spontaneous mutational spectrum of a real organism, *E. coli*. The mutation process included insertions and deletions, but the spectrum of indel lengths was somewhat arbitrary, owing to the paucity of data on spontaneous deletions in *E. coli*. The selection process was modeled by assuming that all frameshift and all nonsense mutations were strongly deleterious. The probability of accepting a nonsynonymous mutation, i.e., the dN/dS ratio, was set to 0.1, a value that is similar to the dN/dS ratio estimated for the Drosophila *adh* gene (Yang et al. 2000). The probability of accepting an indel mutation was set to 0.1 (tables 1 and 4) or to 0.025 (table 5). Those values are somewhat arbitrary as we have essentially no information on the distribution of the selective consequences of indel mutations. Simulations of 64 taxa under those conditions result in true trees that resemble deep phylogenies based on similar numbers of taxa (see fig. 2).

This approach to the simulation of sequence evolution requires, as does reality, that the sequences are aligned with a multiple alignment program such as ClustalX. The average Q-scores from alignment of the protein sequences are quite similar to those from actual deep-phylogeny data sets. Average Q-scores for phylogenies of some antibiotic resistance genes, where the roots of those phylogenies are older than about one billion years ago, include: the class A β-lactamases (Hall and Barlow 2004), average Q-score = 18.6; the class D β-lactamases (Barlow and Hall 2002), average Q-score = 22.8, the class B1 + B2 metallo-β-lactamases (Hall, Salipante, and Barlow 2004), average Q-score = 14.8, and the class B3 metallo-β-lactamases (Hall, Salipante, and Barlow 2004), average Q-score = 11.9. The similarities of the average Q-scores of the simulated tree sequence alignments and the average Q-scores of real alignments supports the contention that the simulations are biologically realistic.

The program ROSE (Stoye, Evers, and Meyer 1998), which also simulates sequence evolution with indels, deserves some comment. Like EvolveAGene, ROSE can be initiated with a root sequence chosen by the user. ROSE evolves DNA sequences according to one of several models

Table 5
Reproducibility of Tree Scores and Topology Scores When the Probability of Accepting an Indel Is Reduced

Replicate	Mean Protein Alignment Q- Score	NJ Protein	Parsimony Protein	Bayesian Protein	NJ DNA-CA	NJ K2P[b] DNA-CA	Parsimony DNA-CA	Weighted Parsimony Tv = 2 DNA-CA	ML DNA -CA	Bayesian DNA-CA
Tree scores										
1	19.6	0.602	0.760	0.856	0.401	0.579	0.502	0.686	0.801	0.834
2	21.1	0.614	0.778	0.848	0.424	0.605	0.502	0.708	0.809	0.839
3	19.6	0.584	0.767	0.805	0.411	0.575	0.499	0.677	0.776	0.791
4	19.6	0.549	0.769	0.844	0.541	0.541	0.473	0.662	0.815	0.845
5	20.0	0.608	0.751	0.809	0.415	0.601	0.481	0.691	0.806	0.828
Mean ± S.E.	19.99 ± 0.30	0.591 ± 0.012	0.765 ± 0.005	0.832 ± 0.011	0.438 ± 0.026	0.580 ± 0.011	0.491 ± 0.006	0.685 ± 0.008	0.801 ± 0.007	0.827 ± 0.010
Topology scores										
1	19.6	0.967	0.967	0.967	0.968	0.952	0.976	0.960	0.952	0.968
2	21.1	0.968	0.984	0.984	0.984	1.000	0.968	0.968	0.968	0.984
3	19.6	0.968	0.984	0.984	1.000	0.968	0.979	0.976	0.968	0.968
4	19.6	0.936	0.968	0.984	0.952	0.952	0.944	0.952	0.968	0.984
5	20.0	0.936	0.934	0.952	0.952	0.968	0.927	0.936	0.952	0.968
Mean ± S.E.	19.99 ± 0.30	0.955 ± 0.008	0.967 ± 0.009	0.974 ± 0.006	0.971 ± 0.009	0.968 ± 0.009	0.959 ± 0.010	0.958 ± 0.007	0.962 ± 0.004	0.974 ± 0.004

[a] Sixty-four taxa, branch lengths were an average of 250 accepted mutations. Probabilities of accepting nonsynonymous substitutions were 0.1; probabilities of accepting insertions and deletions were 0.025.
[b] K2P: Kimura two-parameter model of nucleotide substitution.

(JC [Jukes and Cantor 1969], K2P [Kimura 1980], HKY [Hasegawa, Kishino, and Yano 1985], or F81 [Felsenstein 1981]) while including insertions or deletions at rates specified by the user. Unlike EvolveAGene, ROSE does not separate the processes of mutation and selection, nor does it use an experimentally determined mutational spectrum to set the probabilities of the different types of mutation. One of the primary motivations for the present study was to compare the relative accuracies of trees based on DNA coding sequences and trees based on their corresponding protein sequences. Those comparisons require translating the simulated DNA coding sequences in order to obtain the corresponding protein sequences. Because ROSE can also use a protein sequence as its root, and because it can directly evolve that protein sequence, ROSE does not exclude either frameshifts or nonsense mutations during the process of evolving DNA sequences. As a result translation of frameshifted ROSE sequences produces meaningless protein sequences that do not align. In reality, of course, such frameshifted genes are eliminated by purifying selection. Because ROSE allows frameshifted sequences to persist it was not suitable for use in this study.

One of the questions this study was designed to address was whether alignment of coding sequences according to alignment of the corresponding protein sequences produces more accurate trees than does direct alignment of the DNA sequences themselves. On the basis of paired *t*-tests, the data in tables 1 and 2 make it clear that direct alignment of coding sequences is less preferable ($P < 0.001$).

Another question this study was designed to address was whether protein sequences or DNA coding sequences gave more accurate trees. The tree scores from table 5 were pooled with their corresponding tree scores from table 3, and the topology scores from table 5 were pooled with their corresponding topology scores from table 4, and each of the protein-based methods was compared with its corresponding DNA-CA–based method (NJ with NJ-K2P, Parsimony with Weighted Parsimony, and Bayesian with Bayesian) by paired *t*-tests. When Neighbor Joining was used, the protein tree scores were not significantly different from the DNA-CA tree scores, but the topology scores of the NJ K2P DNA-CA were significantly higher ($P = 0.005$) than those of the NJ protein trees. When parsimony was used, protein tree scores were significantly higher than Weighted Parsimony for DNA-CA ($P = 0.005$), but topology scores were not significantly different. When the Bayesian method was used, DNA-CA tree scores were significantly higher than protein tree scores ($P = 0.01$), but topology scores were not significantly different. These results do not permit making any generalized statement about the relative accuracies of protein-based versus DNA-CA–based phylogenetic trees.

When paired *t*-tests were used to compare all of the methods for the pooled data, Bayesian DNA-CA trees had significantly higher tree scores ($P < 0.0001$) and topology scores ($P < 0.0001$) than any other trees, followed by ML DNA-CA and Parsimony protein trees. Based on accuracy alone, it would appear that the method of choice would be Bayesian estimation of trees based on CodonAligned DNA sequences. There is, however, the matter of the time required to perform the analyses. Table 6 shows the times

Table 6
Time Comparisons

Phylogenetic Method	Time Required, seconds (hours : minutes)
NJ protein	<0.01
NJ DNA-CA	<0.01
NJ K2P DNA-CA	<0.01
Parsimony protein	2.6
Parsimony DNA-CA	2.7
Weighted parsimony DNA-CA	313 (0:5.2)
ML DNA-CA	24,130 (6:42)
Bayesian DNA-CA	27,754 (7:43)
Bayesian protein	54,914 (15:16)

Determined on a Macintosh G4 1.42 GHz computer using OS10.3 for a set of 64 taxa simulated with probabilities of accepting nonsynonymous base substitutions, insertions, and deletions all = 0.1 and an average branch length of 250 accepted mutations.

required for each method to estimate a tree from the same data set. It requires over 10,000 times as much time to estimate DNA-CA Bayesian trees as it does to estimate protein-based parsimony trees. Based on the pooled data, that time increase results in an 8.2% increase in tree score, and a 2.5% increase in topology score. However, the Bayesian method, as implemented by MrBayes, includes the estimation of branch support as posterior probabilities within the time required for the run. Comparable branch support for the protein parsimony requires running at least 2,000 bootstrap replications, which reduces the time advantage for protein parsimony to a mere factor of 5.3.

Although it performs very well when ClustalX alignment quality scores are ⩾20, the Bayesian estimation of protein trees is particularly sensitive to the quality of the alignment, and it is the slowest of the methods tested. Both factors argue against it as a method of choice.

The most reasonable choice might well be to use both parsimony estimation of protein trees and Bayesian estimation of CodonAligned DNA trees, and to point out that any topological differences represent real uncertainty.

A Final Note

EvolveAGene permits accepted mutations, both amino acid substitutions and indels, to occur randomly over the length of the gene. In reality, selective constraints vary in different regions of proteins; e.g., indels, and to a lesser extent amino acid substitutions, tend to appear more often in surface loops than in helices within proteins. That tendency produces runs or blocks within which few gaps appear in alignments. The realism of EvolveAGene could be improved by permitting the user to specify different selective constraints within different regions of the input (root) sequence. Similarly, not all amino acid substitutions are equally likely; e.g., leucine is more likely to be replaced by valine than it is by proline, even though each requires a single nucleotide change. The realism of EvolveAGene could be improved by permitting the user to include a substitution matrix, such as a PAM matrix, that would modify the probability of accepting an amino acid replacement mutation. At present, EvolveAGene permits only purifying selection. In reality, both the intensity and direction of selection vary over the branches of a tree. It would increase the realism of the simulation if the user could apply different intensities and directions of selection over various portions of the tree. The approach to the simulation of sequence evolution exemplified by EvolveAGene should therefore be considered not an end, but a beginning of a more biological approach to sequence evolution simulation.

Acknowledgments

I am grateful to John Huelsenbeck for suggesting the criterion for topology scores. I am grateful to an anonymous reviewer for pointing out the existence of ROSE.

Literature Cited

Barlow, M., and B. G. Hall. 2002. Phylogenetic analysis shows that the oxa β-lactamase genes have been on plasmids for millions of years. J. Mol. Evol. **55**:314–321.

Bull, J. J., M. R. Badgett, H. A. Wichman, J. P. Huelsenbeck, D. M. Hillis, A. Gulati, C. Ho, and I. J. Molineux. 1997. Exceptional convergent evolution in a virus. Genetics **147**:1497–1507.

Doolittle, R. F. 1981. Similar amino acid sequences: chance or common ancestry? Science **214**:149–159.

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

Glickman, B. W., P. A. Burns, and D. F. Fix. 1986. Mechanisms of spontaneous mutagenesis: clues from altered mutational specificity in DNA repair-defective strains. Pp. 259–281 *in* D. M. Shankel, P. E. Hartman, T. Kada, and A. Hollender, eds. Antimutagenesis and anticarcinogenesis mechanisms. Plenum Press, New York.

Hall, B. G. 1999. The spectra of spontaneous growth-dependent and adaptive mutations in ebgR. J. Bacteriol. **181**:1149–1155.

Hall, B. G. 2004*a*. CodonAlign. Macintosh, Windows, Unix Distributed by the author., Bellingham, WA.

Hall, B. G. 2004*b*. Phylogenetic Trees Made Easy: A How-To Manual. Sinauer Associates, Sunderland, Mass.

Hall, B. G. 2004*c*. TuneClustalX. Macintosh Distributed by the author, Bellingham, WA.

Hall, B. G. 2004*d*. CompareTrees. Macintosh Distributed by the author, Bellingham, WA.

Hall, B. G. 2004*e*. EvolveAGene. Macintosh Distributed by the author, Bellingham, WA.

Hall, B. G., and M. Barlow. 2004. Evolution of the serine β-lactamases: past, present and future. Drug Resist. Update **7**:111–123.

Hall, B. G., S. J. Salipante, and M. Barlow. 2004. Independent origins of the Subgroup (B1+B2) and the Subgroup B3 metallo-β-lactamases. J. Mol. Evol. **59**:132–140.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science **255**:589–592.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science **264**:671–677.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics **17**:754–755.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21-32 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

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

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

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Mass.

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.

Schaaper, R. M., B. N. Danforth, and B. W. Glickman. 1986. Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the Escherichia coli lacI gene. J. Mol. Biol. **189**:273–284.

Schaaper, R. M., and R. L. Dunn. 1991. Spontaneous mutations in the Escherichia coli lacI gene. Genetics **129**:317–326.

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

Swofford, D. L. 2000. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. **27**:2682–2690.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Dan Graur, Associate Editor

Accepted December 6, 2004