Syst. Biol. 54(6):900–915, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500354829

Calculating the Evolutionary Rates of Different Genes: A Fast, Accurate Estimator with Applications to Maximum Likelihood Phylogenetic Analysis

RACHEL B. BEVAN,[1] B. FRANZ LANG,[2] AND DAVID BRYANT[1]

[1]*McGill Centre for Bioinformatics, Duff Medical Building, 3775 University Street, Montréal, Quebec, H3A 2B4, Canada*
[2]*Program in Evolutionary Biology, Canadian Institute for Advanced Research; Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada;*
E-mail: rachel@mcb.mcgill.ca. (R.B.B.)

Abstract.—In phylogenetic analyses with combined multigene or multiprotein data sets, accounting for differing evolutionary dynamics at different loci is essential for accurate tree prediction. Existing maximum likelihood (ML) and Bayesian approaches are computationally intensive. We present an alternative approach that is orders of magnitude faster. The method, Distance Rates (DistR), estimates rates based upon distances derived from gene/protein sequence data. Simulation studies indicate that this technique is accurate compared with other methods and robust to missing sequence data. The DistR method was applied to a fungal mitochondrial data set, and the rate estimates compared well to those obtained using existing ML and Bayesian approaches. Inclusion of the protein rates estimated from the DistR method into the ML calculation of trees as a branch length multiplier resulted in a significantly improved fit as measured by the Akaike Information Criterion (AIC). Furthermore, bootstrap support for the ML topology was significantly greater when protein rates were used, and some evident errors in the concatenated ML tree topology (i.e., without protein rates) were corrected. [Bayesian credible intervals; DistR method; multigene phylogeny; PHYML; rate heterogeneity.]

It is widely recognized that the analysis of multiple unlinked genes is superior to single gene analyses for phylogenetic reconstruction. These unlinked genes may, however, be evolving according to very different rules. Heterogeneity of the evolutionary process must be accounted for in phylogenetic analyses (Bapteste et al., 2002; Bull et al., 1993; Huelsenbeck et al., 1996; Nylander et al., 2004; Pupko et al., 2002b; Yang, 1996). The concept of accounting for differing evolutionary pressures within phylogenetic analysis is not new (Yang, 1993). Site-specific rates of evolution can be computed for amino acids (e.g., Rate4Site, Mayrose et al., 2004; Pupko et al., 2002a) and DNA (e.g., DNArates, Olsen et al., 1993) using both Bayesian and maximum likelihood approaches.

Site rates within a gene are likely to be more correlated than rates for sites in different genes. To account for this, it can be assumed that each gene evolves at a different average rate and that these gene rates are drawn from some common distribution (Cranston and Ranala, 2005; Felsenstein, 2001, 2004a). Both Bayesian (Huelsenbeck and Ronquist, 2001) and maximum likelihood (Pupko et al., 2002b; Yang, 1996) methods exist to estimate gene rates (or more generally, locus rates) but these are computationally expensive.

We present a fast, accurate method to estimate the relative evolutionary rates of genes/proteins. For example, when run on a data set with 63 proteins over 123 taxa the algorithm takes less than a second. The method can be applied to protein or nucleotide data, though here we focus on protein sequences. The basic idea is to use pairwise estimates of evolutionary divergence (distances) to deduce the relative rates of different proteins, even when the proteins are not all present in all of the taxa. Although this approach does not give the ML estimates for the rates (Pupko et al., 2002b, Yang, 1996), it does provide an excellent approximation.

After computing rates they are incorporated as extra parameters into the ML tree search, resulting in improved fit as measured by the AIC. The rates estimated using the DistR procedure have been coded into PHYML version 2.2, available at http://atgc.lirmm.fr/phyml/ (Guindon and Gascuel, 2003). PHYML was used because incorporation of the rates was straightforward and because PHYML is an especially fast implementation of ML.

METHODS

The DistR Method

To begin with, the method will be explained through an example. Figure 1 represents three different protein alignments. Not all taxa are present in all three alignments. Suppose that the three proteins have rates r_1, r_2, and r_3. These rates will affect distances inferred from the alignments. Reversing the problem involves using the pairwise distances between species to estimate the different rates r_1, r_2, and r_3.

Figure 1 outlines two ways of obtaining distances from each protein. In the first method ML trees are constructed and the length of the path between two taxa in these trees is measured (referred to hereafter as *patristic ML distances*). In the second method distances are estimated directly from the alignments, as is customary in distance-based methods (referred to hereafter as *pairwise ML distances*). The end result from both methods is a distance matrix for each protein.

If the rate in one protein is twice the rate in a second protein, then the expected distance estimates from the first protein should be twice the expected distance estimates from the second protein. This should hold, approximately, for both pairwise ML distances and patristic ML distances. Equivalently, the distance estimate from the first protein, divided by two, should be approximately the distance estimate of the second protein.

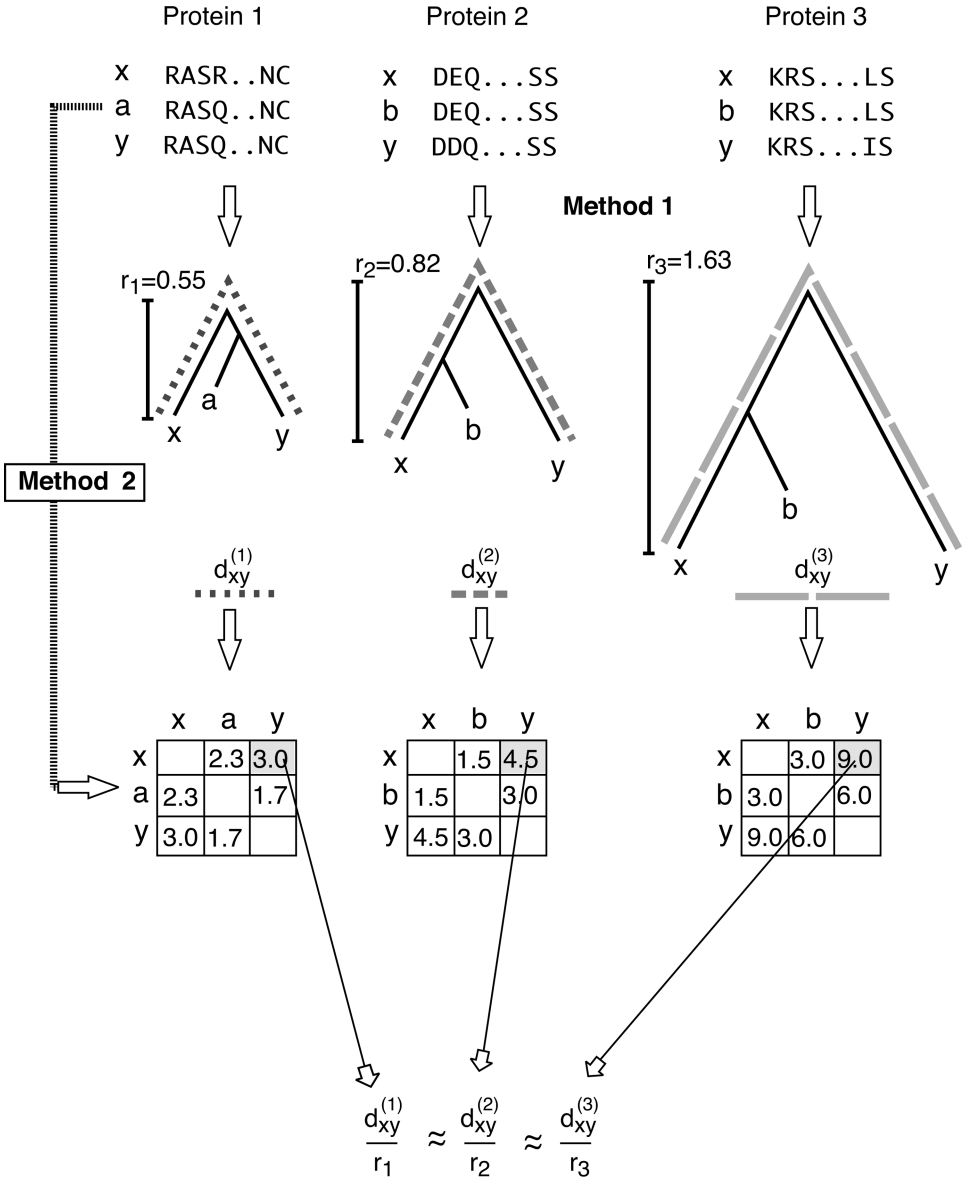

FIGURE 1. The general idea of the DistR estimation procedure. Beginning with individual protein alignments over a set of taxa (with missing data), distances between the species are estimated for each protein alignment. There are two choices of how to estimate the distances: directly from the alignment data (method 2); as the sum of the pairwise distances between taxa on a tree built from the alignment data (method 1). The result is a matrix of pairwise distances between taxa. The ratio of the pairwise distances to the rate of evolution of the protein should be approximately the same for all proteins.

In the example (Fig. 1), and later on, the distance between taxa x and y estimated from protein k is denoted $d_{xy}^{(k)}$, irrespective of whether it is a pairwise or patristic ML distance. Suppose that, for each k, the rate in protein k equals r_k. It follows that $\frac{d_{xy}^{(1)}}{r_1}$ will be approximately equal to $\frac{d_{xy}^{(2)}}{r_2}$ which in turn will be approximately equal to $\frac{d_{xy}^{(3)}}{r_3}$. This is denoted as

$$\frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}, \tag{1}$$

where "≈" means "approximately equal." In Figure 1, this gives $\frac{3.0}{0.55} \approx \frac{4.5}{0.82} \approx \frac{9.0}{1.63}$.

In a sense, the distance estimates obtained from each gene are normalized so that the scale is the same. Define this normalized distance or *consensus distance* between any two taxa as p_{xy}, with the assumption that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}.$$

Assume that rates r_1, r_2, and r_3 in Figure 1 are unknown, whereas the distances remain known. The above

approximate equality leads to

$$p_{xy} \approx \frac{3.0}{r_1} \approx \frac{4.5}{r_2} \approx \frac{9.0}{r_3}. \quad (2)$$

The unknowns p_{xy}, r_1, r_2, and r_3 can be solved for using a least squares approach.

The relation in Equation (2) provides a framework to solve for the *relative rates* r_1, r_2, and r_3, given estimates for the distances $d_{xy}^{(k)}$. This is the basic idea behind the method. The main issues are how to (a) handle the fact that the relations are only approximate; (b) deal with missing distances; (c) compute the rate estimates quickly. These issues are addressed in the following text and in Appendix 2.

To formalize the problem, suppose that there are n proteins (or genes, etc.) over m species. The distance between species x and y derived from protein k is denoted $d_{xy}^{(k)}$. The basic assumption made is that the ratio of the estimated distance between a pair of taxa for a given protein ($d_{xy}^{(k)}$ for protein k and taxa x,y), to the rate of the protein (r_k for protein k), is approximately equal across all proteins.

The rates $r_1, r_2, \ldots, r_n$ are unknown quantities to be estimated based upon the distance data from a given protein alignment. To do this, assume that there exists an unknown consensus distance p_{xy} such that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \ldots \approx \frac{d_{xy}^{(n)}}{r_n},$$

where $n = 3$ for the example in Figure 1. All the consensus distances and rates can now be estimated using a least-squares approach.

In the least squares method it is possible to incorporate measures of uncertainty about the estimated distances $d_{xy}^{(k)}$. Distance estimates with low variance should contribute more to the analysis, whereas distance estimates with high variance (or infinite variance in the case of missing entries) should contribute little. Let $w_{xy}^{(k)} \geq 0$ be a measure of the uncertainty in the distance estimate between taxa x and y derived from protein k. If $d_{xy}^{(k)}$ is accurate, then $w_{xy}^{(k)}$ should be high. If there is less certainty about the accuracy of $d_{xy}^{(k)}$, then $w_{xy}^{(k)}$ should be low. This is achieved using the inverse of the variance of $d_{xy}^{(k)}$, that is, $w_{xy}^{(k)} = \frac{1}{\mathrm{Var}(d_{xy}^{(k)})}$. If protein k is not present in both x and y, then $w_{xy}^{(k)} = 0$. To measure the variance of the distance estimates the approximate formula of Bulmer (1991) is used in the implementation of DistR. Other variance estimators could also be used.

Under a weighted least-squares (WLS) framework the total discrepancy between the ratios $\frac{d_{xy}^{(n)}}{r_n}$ and the consensus distances p_{xy} is measured by

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k} \right)^2 \quad (3)$$

where $\mathbf{p}$ denotes the vector $[p_{12}, p_{13}, \ldots, p_{\frac{(m-1)m}{2}}]^T$ and $\mathbf{r}$ denotes the vector $[r_1, \ldots, r_n]^T$. This is similar to the minimization function used by Lapointe and Cucumel (1997) in the average consensus method. The main difference is that they assume one rate over all proteins, whereas this method includes different rates for each protein. Note that if taxa x and y are missing from a protein k then an estimate for $d_{xy}^{(k)}$ cannot be obtained. However, this is not a problem since the weight $w_{xy}^{(k)}$ will be zero in this case.

Estimating both rates and consensus distances using $q(\mathbf{p}, \mathbf{r})$ leads to the problem of *nonidentifiability*. In the absence of any error each estimated protein distance $d_{xy}^{(k)}$ is the product of the rate of the protein r_k and the consensus distance p_{xy}. Thus, a perfect fit to the equation is still achieved if all the rates are multiplied by some constant and all the consensus distances divided by the same constant. There is a problem of determining scale. Hence, Equation (3) does not have a well-defined minimum. To solve this problem a constraint

$$\sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} p_{xy} = \kappa \quad (4)$$

must be added to system, where κ is an arbitrary positive constant. The particular value of κ is irrelevant since changing κ merely causes all estimated rates to be multiplied by the same constant value. For this reason, it is possible to infer *relative rates* only. In DistR $\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$, thus constraining the weighted estimated distances to be equal to the weighted consensus distances. This was empirically determined to minimize the variance of the DistR estimates.

Appendix 3 describes an extremely fast algorithm for minimizing the function $q(\mathbf{p}, \mathbf{r})$ subject to the constraint in Equation (4). The algorithm takes $O(nm^2 + n^3)$ time and $O(n^2 + m^2)$ memory. For example, when run on a data set with 63 proteins over 123 taxa, the algorithm takes less than a second. An implementation with source code is available at http://www.mcb.mcgill.ca/~rachel.

Experimental Studies

An extremely rapid method for estimating the relative rates of different genes has been proposed. The method is orders of magnitude faster than existing ML and Bayesian approaches. The most important question remaining is to what extent this increase in speed affects the accuracy of the estimates. In order to address this question, the accuracy of the new method was assessed using both simulated and empirical data.

In all the analyses PHYML (version 2.2) was used (Guindon and Gascuel, 2003) to compute ML distances and trees, with a JTT protein model, eight gamma categories plus invariant sites and the default (BIONJ) starting tree. The gamma shape parameter and proportion of invariant sites were estimated using default optimization routines in the program. When constructing ML trees from real data several bootstrap values were computed. As detailed below these values depend upon: whether patristic or pairwise ML distances were used in the DistR procedure; whether the rates were reestimated for each bootstrap replicate.

For both the simulated and empirical data, DistR estimates based upon patristic and ML distances were compared. This comparison was made in order to determine whether or not the additional computational effort required for estimating patristic ML distances is justified.

Experimental Studies—Simulated Data

The two key questions addressed through the simulation studies are:

- *Patristic versus pairwise ML distances*.—How accurate are the rate estimates using pairwise versus patristic ML distances?
- *Missing distances between taxa*.—How are DistR rate estimates affected when proteins are not present in all taxa?

To answer these questions protein alignments were simulated using Pseq-Gen (Grassly et al., 1997) with the JTT model of evolution. The initial tree and branch lengths were taken from an independent analysis of mitochondrial Atp8 proteins in 58 eukaryotes. Two types of simulations were carried out. The first, intended to address the first question, involved construction of 20 protein trees by randomly deleting taxa from the starting tree. In total there were four protein trees with 53 taxa, four with 48 taxa, four with 43 taxa, four with 38 taxa, and four with 33 taxa. For each tree a rate was sampled from a precomputed distribution of rates based on real data (data not shown), and protein alignments of length 100, 300, 500, and 1000 generated using Pseq-Gen (Grassly et al., 1997) (note that the average length of naturally occurring proteins is approximately 300 amino acids). The second analysis, intended to address the second question, increased the number of taxa deleted from the starting tree. In total there were seven trees with 25% of the taxa, seven with 50% of the taxa, and seven with 75% of the taxa. This resulted in 21 trees, 7 each with 16, 30, and 44 taxa, respectively. For each tree a rate was sampled from a precomputed distribution of rates based on real data (data not shown), and protein alignments of length 1000 generated using Pseq-Gen (Grassly et al., 1997). This experiment follows a protocol proposed by (Eulenstein et al., 2004). For both experiments, and for every set of parameters, 10 replicates of the experiment were performed. See Figure 2 for an overview of the simulations.

Statistics measured on the simulated data, including goodness-of-fit and mean squared error, are explained in detail in Appendix 1. These statistics were used to relate the accuracy of the DistR rate estimates to the known rates at which the proteins were simulated.

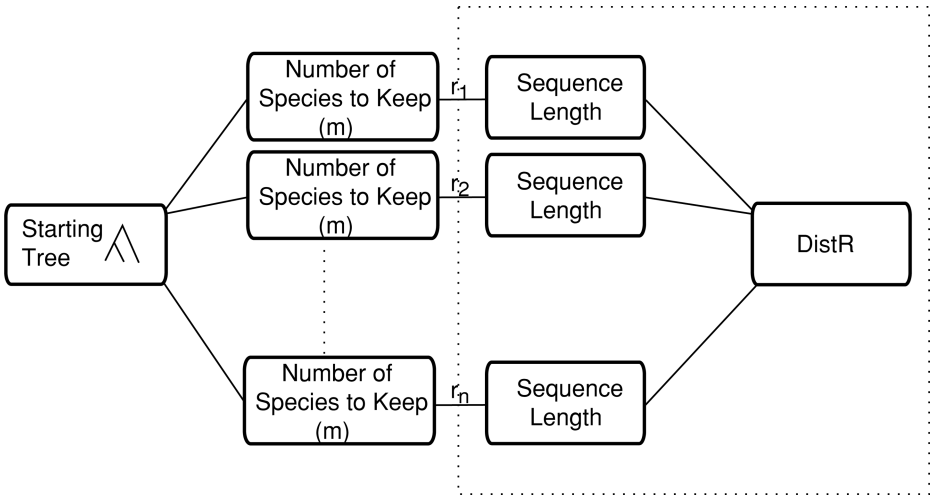

FIGURE 2. The general flow of the simulation studies. Two studies were performed, one with $n = 20$ and the other with $n = 21$ (where n is the number or proteins). The first study compared different methods of estimating distances using different alignment lengths. In the first study, 20 random subtrees from an original tree of 58 species were created, four each of size $m = 33$, $m = 38$, $m = 43$, $m = 48$, and $m = 53$ (where m is the size of the taxon set for a given protein). For each tree, a rate was sampled from a precomputed distribution of rates based on real data (data not shown). Protein alignments of length 100, 300, 500, and 1000 were simulated using Pseq-Gen (Grassly et al., 1997). A second analysis compared rate estimates with increasing amounts of data. Twenty-one random subtrees from the original tree of 58 species were created, 7 each of size $m = 16$, $m = 30$, and $m = 44$ (corresponding to approximately 25%, 50%, and 75% of the species [as in Eulenstein et al., 2004]). For each tree, a rate was sampled from a precomputed distribution of rates based on real data (data not shown). Alignments of length 1000 were generated. For both studies, 10 replicates were performed for each set of parameters.

Experimental Studies—Empirical Data

The data analyzed in this study consist of a set of 15 aligned mitochondrial protein sequences from 29 taxa. The taxon names and accession numbers are given in Table 1. Protein names and alignment accession numbers appear in Table 2. This multiprotein data set is of moderate size, and variants thereof have been used in numerous publications (e.g., Bullerwell et al., 2003; Lang et al., 2002; Sumida et al., 2001; Tomita et al., 2002). Furthermore, some of the species have high evolutionary rates and substitutional saturation of sites (i.e., *Smittium*), whereas others have very short branches in the resulting phylogenetic tree. Combined, these two properties can cause inaccurate grouping of the taxa due to long-branch attraction artifacts (Felsenstein, 1978).

TABLE 1. Empirical data analyzed. Names and accession numbers for protein sequences studied from Fungal species and outgroup. Fifteen proteins were downloaded for each species (if present in the species), the names of which are in Table 2.

Species	GenBank accession number
Ascomycota	
Aspergillus nidulans	CAA33481, AAA99207, AAA31737, CAA25707, AAA31736, CAA23994, X15442, P15956, CAA23995, CAA33116, X00790, X15441, X06960, J01387, X01507
Candida albicans	AF285261
Candida glabrata	CGL511533
Hypocrea jecorina	AF447590
Penicillium marneffei	NC_005256
Pichia canadensis	NC_001762
Podospora anserina	X55026
Saccharomyces cerevisiae	AJ_011856
Schizosaccharomyces japonicus	NC_004332
Schizosaccharomyces octosporus	AF275271
Schizosaccharomyces pombe	X54421
Torrubiella confragosa	AF487277
Yarrowia lipolytica	AJ307410
Basidiomycota	
Cryptococcus neoformans	NC_004336
Schizophyllum commune	AF402141
Cantharellus cibarius [a]	
Choanoflagellida	
Monosiga brevicollis	AF538053
Chytridiomycota	
Allomyces macrogynus	U41288
Harpochytrium94	NC_004760
Harpochytrium105	NC_004623
Hyaloraphidium curvatum	AF402142
Monoblepharella	AY182007
Rhizophydium136	NC_003053
Spizellomyces punctatus	AF402142
Metazoa	
Homo sapiens	NC_001807
Metridium senile	AF000023
Zygomycota	
Smittium culisetae	AY8632133
Mortierella verticillata	AY863211
Rhizopus oryzae	AY863212

[a] Downloaded from http://megasun.bch.umontreal.ca/People/lang/FMGP/proteins.html.

Alignments were performed using the default settings of ClustalW (Thompson et al., 1994). Highly variable sites or those with many gaps were eliminated using Gblocks (Castresana, 2000) with the following settings: number of sequences for a flank position equal to half the number of species plus one; number of contiguous nonconserved positions equal to 10; minimum length of a block four; half the species allowed gaps. All other parameters were set to default.

The key questions addressed using real protein data are:

- *Comparison of DistR estimates to ML estimates*.—How do DistR rate estimates compare to those obtained using the ML based method COMBINE (Pupko et al., 2002b)?
- *Comparison of DistR estimates to Bayesian estimates*.—How do DistR rate estimates compare to those obtained by MrBayes (Huelsenbeck and Ronquist, 2001) under a Bayesian approach?
- *Patristic versus pairwise ML distances*.—How do rate estimates from pairwise ML distances and rate estimates from patristic ML distances compare when applied to real data?
- *Inclusion of DistR estimates into the phylogenetic tree search of PHYML*.—What is the affect of including DistR estimates in an ML tree search? Is there a significantly improved fit? Are improved phylogenetic estimates obtained?

Comparison of DistR estimates to ML estimates.—Note that when comparing DistR rates to those computed using COMBINE (Pupko et al., 2002b), the number of taxa and proteins had to be restricted, because COMBINE can currently only handle data sets for which all taxa are present in all proteins. Two different starting trees were included in the analysis: the ML tree from PHYML based upon the concatenated data set and the ML tree from PHYML when protein rates were incorporated. Rates were estimated under three different models: global amino acid frequencies with one gamma distribution; local amino acid frequencies (for each protein partition) with one gamma distribution; local amino acid frequencies with one gamma distribution for each partition.

Comparison of DistR estimates to Bayesian estimates.—Bayesian estimation of the posterior distribution of the protein rates was performed using MrBayes version 3.0 (Huelsenbeck and Ronquist, 2001). Default priors were used with the JTT model of evolution plus one gamma distribution (eight categories), one parameter for the proportion of invariant sites, and one set of branch lengths for the entire data set. This is the same model that is used for the PHYML + protein rates analysis of the data. Two runs of four chains with 300,000 iterations were performed; the burn-in used was 30,000. A further analysis of the data was performed without protein rates (using the same model) in order to compare to the concatenated PHYML analysis. Four chains were run for 150,000 iterations, with a burn-in of 15,000. Convergence of the chains was determined empirically.

TABLE 2. DistR estimates for empirical data based on pairwise and patristic ML distance estimates. Mean rate estimates and variances for rate estimates are based upon bootstrap replicates over the fungal data set. Rates are normalized so that the average rate is one. Acc. no. = accession number for the alignment in EMBL. AL = alignment length. Patristic refers to rates estimated based on distances from maximum likelihood trees. Pairwise refers to rates estimated based on maximum likelihood distances.

				Patristic		Pairwise	
Protein	Acc. no.	No. of species	AL	Mean	Variance $\times 10^{-3}$	Mean	Variance $\times 10^{-3}$
Atp8	ALIGN_000885	28	32	1.08	8.68	1.15	11.8
Atp9	ALIGN_000886	26	73	0.55	5.12	0.55	4.35
Rps3	ALIGN_000900	11	77	2.02	41.1	2.33	31.5
Nad3	ALIGN_000893	24	79	1.13	8.82	1.15	10.1
Nad4	ALIGN_000894	24	424	1.14	3.52	1.10	2.76
Nad4L	ALIGN_000895	23	85	0.87	5.91	0.91	6.45
Nad6	ALIGN_000897	24	96	1.05	7.21	1.10	7.80
Atp6	ALIGN_000884	29	203	1.07	3.76	1.03	4.07
Cox2	ALIGN_000889	29	220	0.75	3.81	0.71	2.98
Cox3	ALIGN_000890	29	245	1.05	4.75	0.86	3.24
Nad1	ALIGN_000891	24	294	0.89	2.61	0.84	2.30
Nad2	ALIGN_000892	23	313	1.21	2.16	1.29	2.69
Cob	ALIGN_000887	29	375	0.67	1.17	0.61	1.04
Cox1	ALIGN_000888	29	487	0.53	1.76	0.46	.749
Nad5	ALIGN_000896	24	520	1.01	2.79	0.89	1.94

Inclusion of DistR estimates into the phylogenetic tree search of PHYML.—DistR rates were incorporated into the ML framework of PHYML following the proportional approach (Pupko et al., 2002b; Yang, 1996); however, optimization over the rates was not performed. ML trees over the entire data set were calculated in four different ways using this modified version of PHYML. In the first analysis, the proteins were simply concatenated (equivalent to a rate of one for each protein). In the second analysis, the estimated protein rates from the real data set (based on patristic ML distances) were used for each bootstrap replicate when computing the likelihood. In the third and fourth analyses, protein rates were estimated for each bootstrap replicate using patristic and pairwise ML distances respectively. These rates were incorporated into the likelihood computation for each bootstrap replicate. Consensus trees were computed using the CONSENSE program available in the PHYLIP package (Felsenstein, 2004b).

RESULTS AND DISCUSSION

Simulated Data

Patristic versus pairwise ML distances.—The first simulation study demonstrates two important results: pairwise ML distances provide equally good distance estimates as patristic ML distances to the DistR method (Fig. 3); if the fit of the initial pairwise/patristic ML distances to the data is accurate then the DistR estimates will be accurate (Figs. 3 and 4). The first result is important since pairwise ML distances are very fast to compute. The second result indicates that error in the rate estimates stems principally from error in the distance estimates, rather than the DistR method itself.

The numerical results from the first experiment are summarized in Figure 3. The proteins are sorted in order of increasing rate, and the histogram indicates the mean squared error (MSE) over the 10 different replicates (see Appendix 1 for the exact formula used to compute MSE). Mean rate estimates are labelled to the right of each MSE bar, with the rate at which the data was simulated on the left. Results are presented only for alignments of length 100 and 1000. The results for alignments of length 300 and 500 fall in-between these two extremes. Note that the MSE increases in proportion to the rate, so results are presented on two scales.

The mean estimates for the different methods were quite close to the real rates at which the data were simulated, regardless of the alignment length, procedure used to estimate the distances, or rate at which the data was simulated (Fig. 3). However, it is clear from the mean squared error that the DistR estimates based on shorter alignments have larger error (or greater variation), despite the fact that the mean rate estimate is often almost as accurate as that for longer alignments. Furthermore, the mean squared error tends to increase with higher rates. This is likely because the error is often in the third significant digit; for slower rates this will lead to a smaller MSE. Overall there is negligible difference between the mean and MSE statistics for a given alignment length (comparing DistR estimates based on patristic versus pairwise ML distances).

Results also indicate that errors in the rate estimates are due to errors in the original distances rather than approximations introduced in the DistR method. For each protein and alignment length the absolute error between the mean rate estimates and the real rate at which the alignments were simulated was compared to the goodness-of-fit between the estimated and true distances (Fig. 4). This fit can be measured since the data are simulated under a known model at a particular rate. Alignments of length 100 and 300 only were examined, since the errors become negligible for longer alignments. The fit was measured using the goodness-of-fit statistic of Tanaka et al. (Tanaka and Huba, 1985), which is determined from the sum of squares error between true and estimated distances, normalized by the sum of the true distances squared. The exact formula for goodness-of-fit is

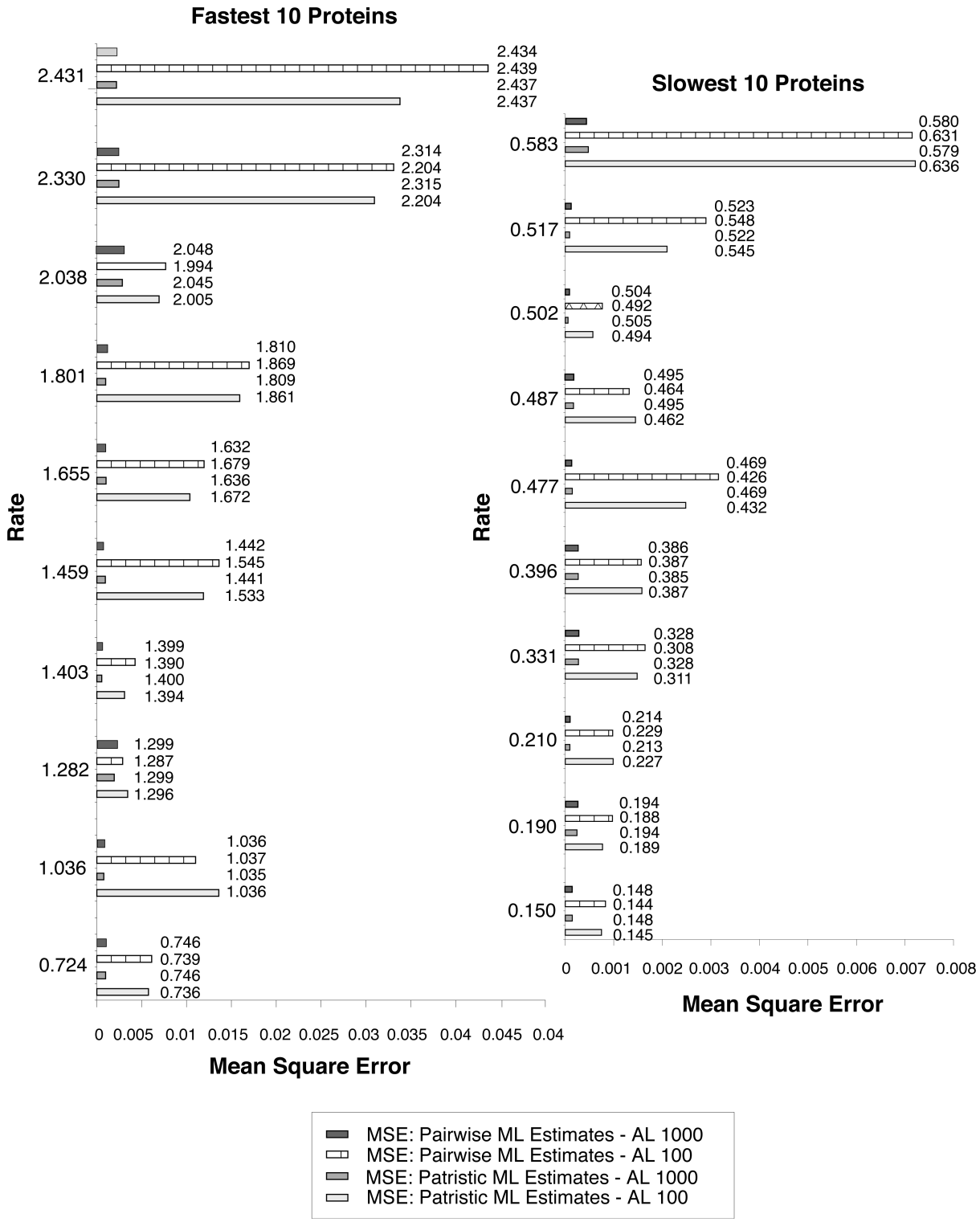

FIGURE 3. Mean squared error for different methods of distance estimation and different alignment lengths. The rates at which the data were simulated are labeled on the left-hand side of the graph. The mean rate estimate for a given distance estimation method, alignment length, and rate is given on the right of the MSE bar. AL = alignment length. The 10 fastest proteins are in the left-hand column. The number of species in each protein (from fastest to slowest) are Protein 1: 53 species; Protein 2: 38 species; Protein 3: 33 species; Protein 4: 53 species; Protein 5: 38 species; Protein 6: 48 species; Protein 7: 53 species; Protein 8: 48 species; Protein 9: 43 species; Protein 10: 33 species. The 10 slowest proteins are in the right-hand column. The number of species in each protein (from fastest to slowest) are Protein 1: 33 species; Protein 2: 48 species; Protein 3: 43 species; Protein 4: 43 species; Protein 5: 48 species; Protein 6: 33 species; Protein 7: 43 species; Protein 8: 53 species; Protein 9: 38 species; Protein 10: 38 species. All rates are normalized so that the average rate is one over all 20 proteins. The total number of taxa in the data set is 58.

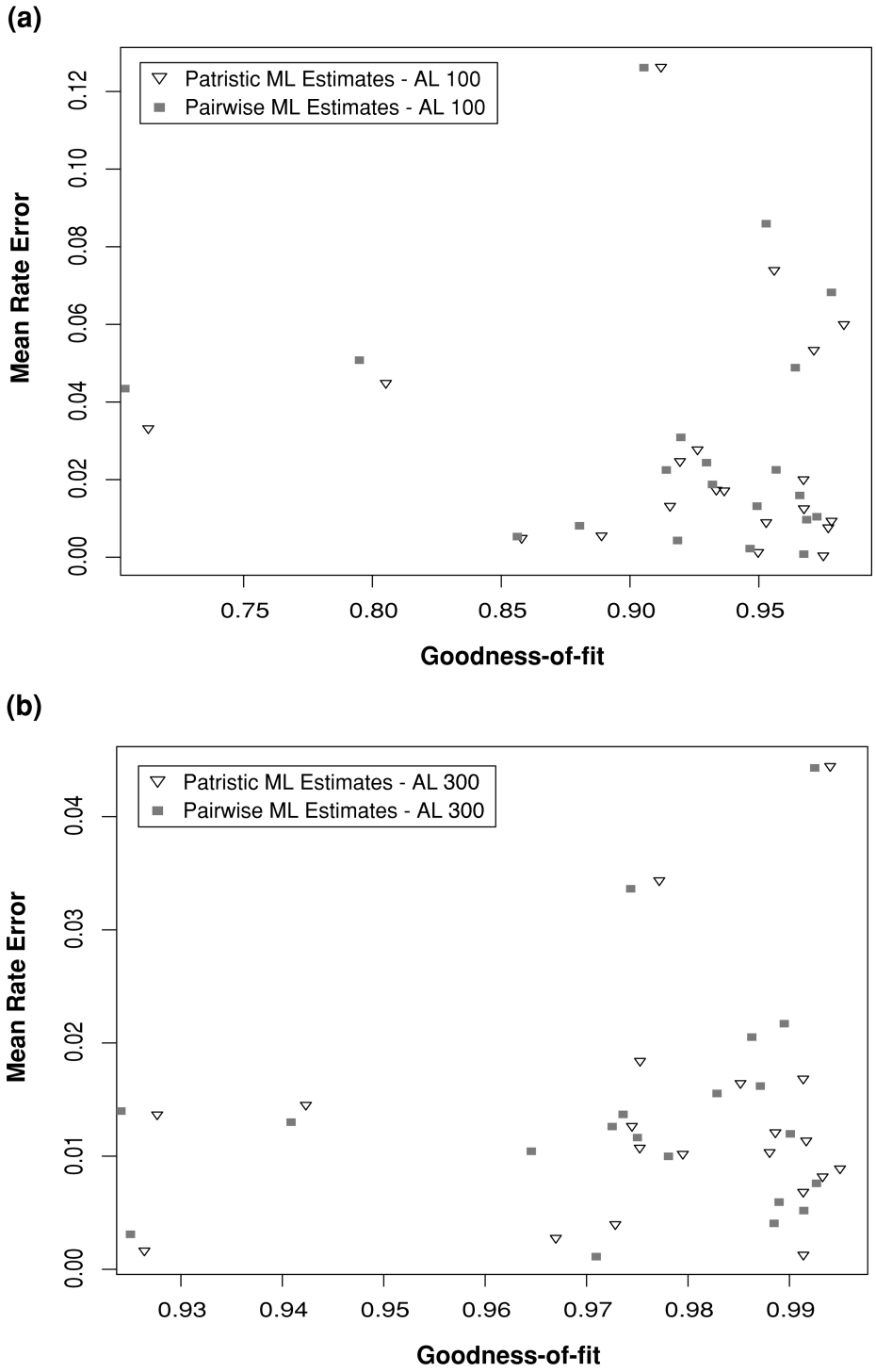

FIGURE 4. Average error of DistR rate estimates compared to goodness-of-fit of distances based upon patristic and pairwise ML distance estimates. (a) DistR rate estimates were based upon simulated proteins of length 100. (b) DistR rate estimates were based upon simulated proteins of length 300. A higher value for goodness-of-fit means that the fit of the estimated distances to the original distances is better.

presented in Appendix 1. The statistic has a maximum of one, which indicates a perfect fit.

It is expected that with longer alignments the goodness-of-fit will increase, indicating that the fit of the model to the data is better. This is clearly the case as seen when comparing goodness-of-fit for alignments of length 100 (Fig. 4a) to that for alignments of length 300 (Fig. 4b). The fit is further improved, and relative error reduced, with alignments of length 500 and longer (data not shown). The decrease in the goodness-of-fit (indicating a worse fit) seen with short alignment lengths indicates that the error of the method is dependent upon

the error of the distance estimates and is not a property of the estimation procedure itself.

Interestingly, the error in rate estimation is in some cases less when based upon pairwise ML distances, rather than patristic ML distances. Given that the multiple sequence alignments are short (100 and 300 amino acid residues) and include many species (at least 33 in each protein alignment), there are many trees that will fit the data equally well. Thus, there is high variation in building a ML tree to fit the original tree on which the data were simulated. Hence, estimating a ML tree with few data will likely lead to an incorrect topology. This will result in a worse fit between the original tree and the tree estimated from the alignment data. This is not true for pairwise ML distances, which do not account for topology.

Missing distances between taxa.—In the previous experiment, less than half of the taxa were missing in each protein, and 20 proteins were used to estimate rates. The effects of more extreme missing taxa were also tested, where no distance estimates were present between some pairs of taxa. To achieve this, up to 75% of the taxa were removed from the starting tree. Additionally, many fewer proteins were used for DistR estimation. Results indicate that the DistR method is robust to missing taxa, though having many missing taxa led to the expected increase in variance of the rate estimates.

Figure 5 summarizes the error in rate estimates for two simulated data sets. In the first example (Fig. 5a) there are four protein trees, each with 16 taxa ($\approx 28\%$ of the total taxon set). In the second example (Fig. 5b) there are eight protein trees. Seven of these have 16 taxa and the other has 30 taxa. The proteins are ordered from fastest to slowest rate in both Figure 5a and Figure 5b. Mean rate estimates are shown on the right of the MSE, and the rate at which the protein simulated (averaged to equal one) is given on the left. Simulated proteins in Figure 5a are labeled from I to IV. The same simulated proteins in Figure 5b are likewise labeled.

Once again it is evident that pairwise ML distances and patristic ML distances give almost identical average relative rate estimates (to within two or three decimal places). Furthermore, the missing data has little effect on mean rate estimates, but does have a large effect on the variance. For instance, comparing the MSE for the first protein in Figure 5a to that of the second protein in Figure 5b (it is the same simulated protein), it is clear that although the mean rate estimate is approximately as accurate with more taxa (Fig. 5b), the MSE is clearly smaller when more distances between a pair of taxa are included in the analysis. Thus it is evident that more data in terms of pairwise distances between taxa (over multiple proteins) will reduce the error of the DistR estimate.

Calculation of the relative rates within groups of the same number of species was also performed (i.e., proteins with 16 species, proteins with 30 species, and proteins with 44 species). For each subset of proteins mean rate estimates based on pairwise ML distances were slightly worse or identical to those based on patristic ML distances (data not shown). In addition, the variances were greater in general for rates estimated based on pairwise ML distances. The major difference between the three analysis was that the variance of the rate estimates was lower when more species were included in the analysis. Furthermore, the mean rate estimates were slightly more accurate for the data sets over larger taxon groups (data not shown).

Accuracy in spite of missing taxa demonstrates that the rate estimation procedure is consistent (assuming that the initial distance estimates are accurate), regardless of the number of proteins under analysis. This is because rates are not computed relative to the distance estimates of one protein. Rather, they are constrained by all the distance estimates. Thus, if one set of distance estimates is extremely biased with respect to the remainder of the distances they will not have a strong effect on the final rate estimates.

Empirical Data

Comparison of DistR estimates to ML estimates.—Rates were calculated in a ML framework using only those proteins that are present over the entire species set (Atp6, Cob, Cox1, Cox2, and Cox3) due to a constraint of the program COMBINE (Pupko et al., 2002b). Table 3 shows the time for rate estimation and rate estimates based on different models under the ML framework in comparison to DistR estimates based on pairwise and patristic ML distances. Two sets of ML estimates are given for each model. The first based upon the concatenated tree, and the second on the DistR incorporated ML tree. DistR estimates are computed far more rapidly and are still accurate in comparison to ML estimates. In comparison to the six ML estimates, the DistR rates based on patristic ML distances are slight overestimates for Cob and Cox1, and slight underestimates for Cox2 and Cox3. The estimate for Atp6 is an average of the 6 ML estimates (Table 3). Notably, the patristic DistR estimates for Cob and Cox1 are closest to the ML estimates based on

TABLE 3. Comparison of ML rate estimates to DistR estimates. Comparison of relative rate estimates and estimation time from COMBINE and DistR for five proteins (Atp6, Cob, Cox1, Cox2, and Cox3) from the fungal data set. For each model, rates based upon the maximum likelihood concatenated tree from PHYML are given on the first line, and rates based upon the maximum likelihood tree incorporating DistR rates (computed in PHYML) are given on the second. All estimates were normalized so that the average rate is one. GF = global amino acid frequencies; LF = local amino acid frequencies (calculated for each protein); 1-GAM = one gamma distribution estimated for the entire data set; 5-GAM = one gamma distribution for each protein; DistR Pat = DistR estimation using patristic ML distances; DistR Pair = DistR estimation using pairwise ML distances.

Method	Time	Atp6	Cob	Cox1	Cox2	Cox3
GF + 1-GAM	776s	1.24	0.81	0.62	0.99	1.34
		1.25	0.81	0.63	0.99	1.33
LF + 1-GAM	842s	1.35	0.80	0.61	0.94	1.31
		1.36	0.80	0.62	0.93	1.30
LF + 5-GAM	648s	1.36	0.79	0.59	0.94	1.31
		1.39	0.78	0.61	0.92	1.30
DistR Pat	0.116s	1.32	0.83	0.66	0.91	1.29
DistR Pair	0.122s	1.40	0.83	0.64	0.96	1.18

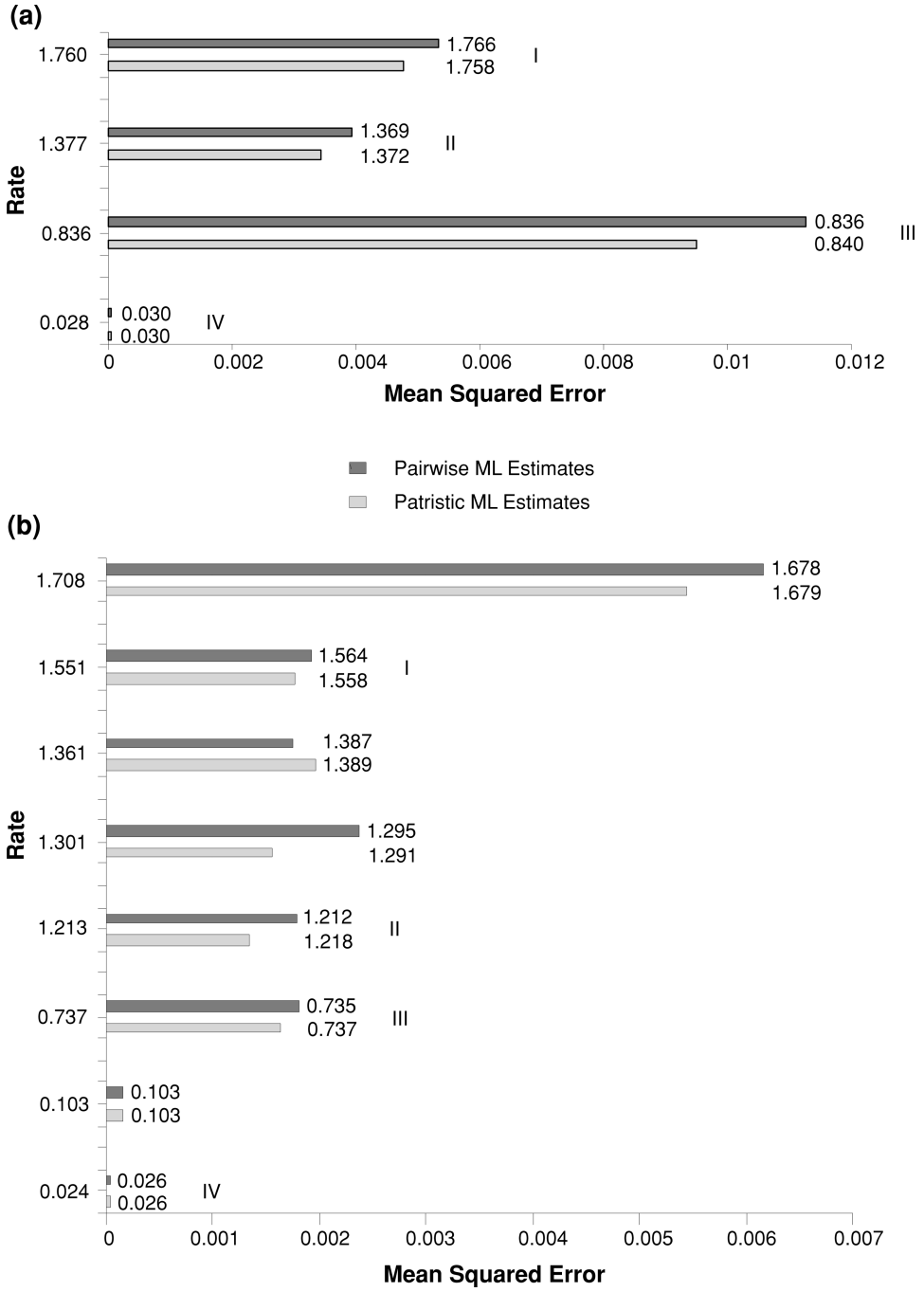

FIGURE 5. Mean squared error for different methods and different amounts of distance data. The rates at which the data were simulated are labelled on the left-hand side of the graph in both (a) and (b). Mean rate estimates for both distance estimation methods are labelled on the right of the MSE bars for each protein. All rates are normalized so that the average rate is one in both (a) and (b) and are sorted from fastest to slowest. Proteins that are the same in both (a) and (b) are labelled. (a) Rate estimates based upon a data set consisting of four proteins with 16 taxa each. (b) Rate estimates based upon a data set consisting of eight proteins; seven with 16 taxa and one with 30 taxa.

the rate-incorporated tree using global amino acid frequencies plus the one-gamma-distribution model. Conversely, the DistR estimates for Cox2 and Cox3 are closest to the ML estimates based on the same tree, using local amino acid frequencies and the five-gamma-distribution model. The DistR estimates based on pairwise ML distances are quite close to those based on patristic ML distances, except for Atp6 and Cox3. Atp6 has a much higher rate—quite close to the ML estimate for the LF + 5-GAM model where the estimates were based on the rate-incorporated ML tree. However, the Cox3 estimate is quite low compared to all ML estimates; Cox3 had a higher variation in rate estimation over all proteins (Table 3), a case where perhaps the lack of topological

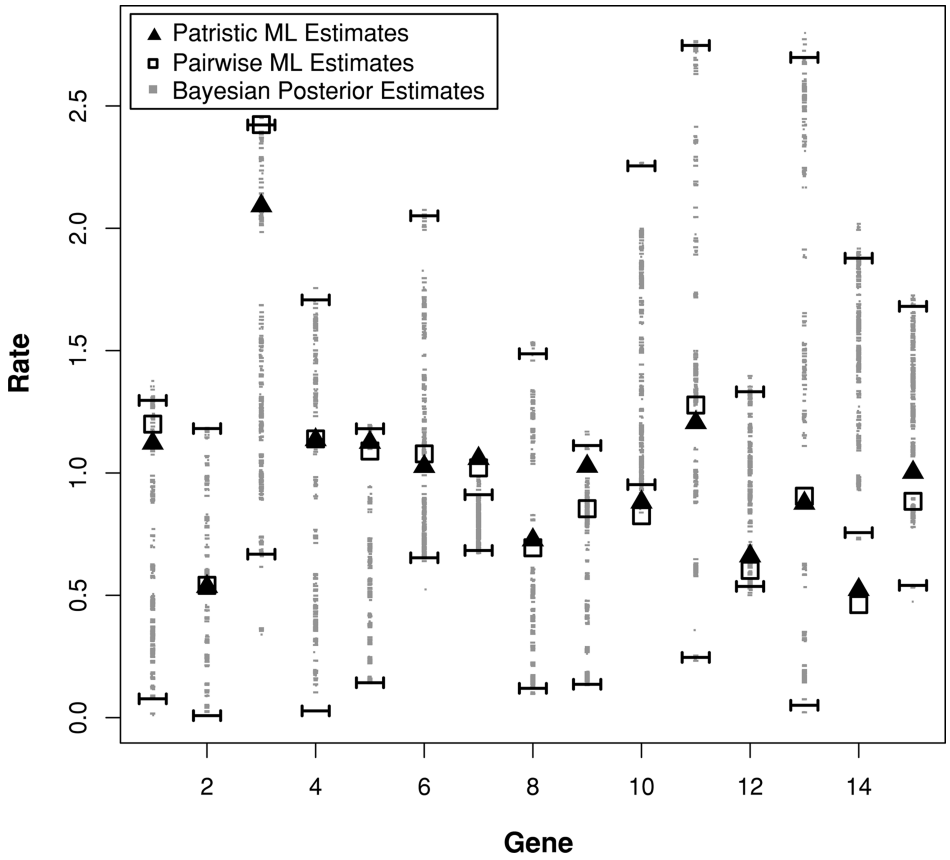

FIGURE 6. Distribution of rates from the MrBayes proportional model analysis compared to DistR estimates. Bars at either end represent the 95% credible interval. The DistR estimate based upon patristic ML distances is marked by a solid triangle. The DistR estimate based upon pairwise ML distances is marked by a square. The posterior rate estimates of MrBayes are given by a solid square. DistR estimates are normalized so that the average rate is one (as in MrBayes). Proteins are ordered from shortest to longest as follows: Atp8, Atp9, Rps3, Nad3, Nad4, Nad4L, Nad6, Atp6, Cox2, Cox3, Nad1, Nad2, Cob, Nad4, Cox1, and Nad5.

information decreases the accuracy of the DistR estimate. Clearly this is not an issue for most proteins, but can be an issue for some. Overall it appears that the DistR estimates are model independent regardless of distance estimation procedure and provide excellent first approximations to the ML estimates.

Comparison of DistR estimates to Bayesian estimates.— The posterior distribution of rates from MrBayes is shown in Figure 6. For all but three of the proteins the DistR estimates fall within the 95% posterior credible interval for the protein rate. Each of Nad6, Cox1, and Cox3 have DistR estimates that do not fall between the 95% posterior credible interval. Both Cox1 and Cox3 have average sequence lengths, and 29 taxa each. Nad6 is shorter at less than 100 amino acids, with only 24 species. In the case of Nad6 perhaps the short sequences length contributes to uncertainty in the DistR estimates. However, it is unlikely that the Bayesian posterior distributions of the rates are accurate. This conclusion is based upon the fact that the four chains were mixing quite poorly in both runs even after 300,000 iterations (data not shown). Sampling from the posterior distribution is unlikely to be correct since the chain might be oversampling from areas of low likelihood. Comparison of the tree of the highest likelihood from this analysis to the tree of highest likelihood based on the concatenated data indicates that MrBayes was in a suboptimal topological space when sampling rate estimates (using the Bayesian information criterion, data not shown). Furthermore, the DistR ML tree is a significantly better fit of the model to the data based on the AIC (Felsenstein, 2004a) when compared to the likelihood of the MrBayes rate incorporated tree as computed in PHYML. Thus, although the posterior distribution of the rates appears reasonable, the chain seems to be having difficulty sampling through topology space.

Thus, it appears that the proportional model under MrBayes, when used without different parameters for each partition (as in Nylander et al., 2004), does not search tree space as well as PHYML with the rate multipliers included. Perhaps this is due to an incorrect prior on the rate parameters used. If this is the problem the DistR method can certainly be used to find a distribution of the rates of proteins, which could be used as the prior on these parameters. The discrepancy could also be due to the different search heuristics used in MrBayes. Given the computational complexity of the search, it might be

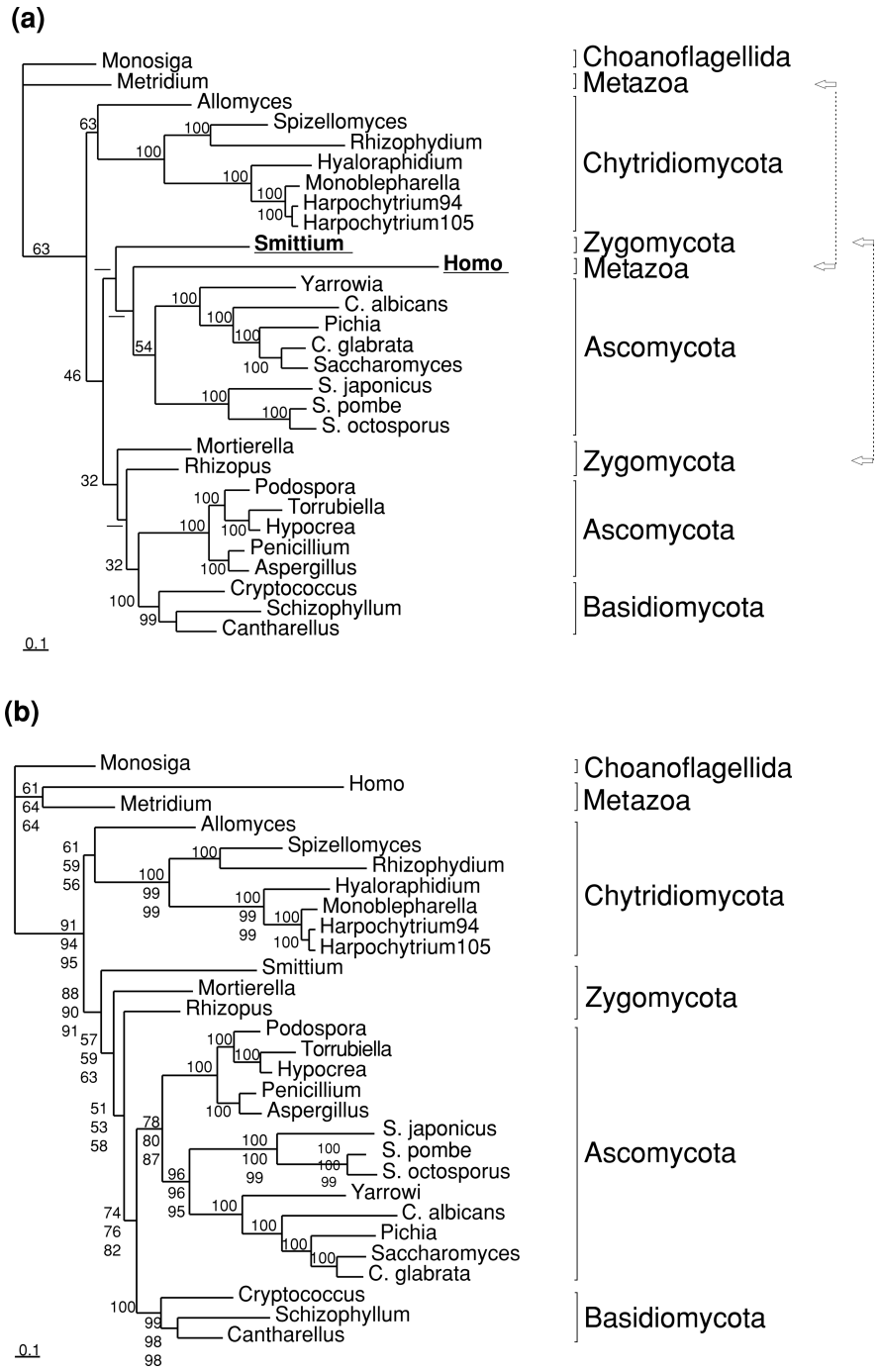

FIGURE 7. (a) Phylogenetic analysis based upon the mitochondrial data set. The topology shown was inferred using PHYML without DistR protein rates, using the JTT model of protein evolution, with eight gamma categories, and ML estimation of the alpha parameter of the gamma distribution and the proportion of invariant sites. It was constructed using the concatenated "unambiguously" aligned proteins. Bootstrap support for this topology was computed based upon 100 replicates. The percentage of support for each clade is given at the root of the clade. In cases where the consensus tree differed from the maximum likelihood topology a "–" is written. (b) Phylogenetic analysis based upon mitochondrial data set. The topology shown was inferred using PHYML with DistR protein rates, using the JTT model of protein evolution, with eight gamma categories, and ML estimation of the alpha parameter of the gamma distribution and the proportion of invariant sites. It was constructed using the concatenated unambiguously aligned proteins and protein rate estimates. The percentage of support for each clade is given. Bootstrap support for this topology was computed based upon 100 replicates, using three different methods. The top numbers give the percentage of support based upon using the patristic ML distance DistR estimates from the real data as rate values in computing the ML tree for each bootstrap replicate. The middle numbers give the percentage of support based upon reestimating DistR estimates for each bootstrap replicate using patristic ML distances. The bottom numbers give the percentage of support based upon reestimating DistR estimates for each bootstrap replicate using pairwise ML distances. When bootstrap support was the same for each method of incorporating rates it is given only once.

difficult for the program to search for the best rate parameters while also searching for the best topology.

Patristic versus pairwise ML distances.—The relative protein rates of the real data are unknown. However the variance of the rate estimates using both patristic and pairwise ML distances can be compared, a smaller estimate being preferable. Contrary to expectations, but confirming the simulation studies, rate estimates from pairwise ML distances had smaller variance than rate estimates from patristic ML distances.

Variances of the rate values computed were estimated by nonparametric bootstrap of the protein alignments, and reestimation of the distances and DistR rates for each bootstrap data set. The mean and variance of the DistR estimates for pairwise and patristic ML distances show some interesting trends (Table 2). In general, the average rate estimates were similar, with the notable exception of Atp8, Cox3, and Rps3 (and to a lesser extent Nad2, Nad5, and Nad6). Ten of the 15 protein rates derived from patristic ML distances had greater variance than their counterparts derived from pairwise ML distances. (Table 2). These results support the conclusion that introducing topology into the distance estimation procedure is not likely to lead to better distances estimates for the DistR procedure when so many taxa are involved and the alignments are short. This is a consequence of the large number of distinct trees that can fit a short alignment equally well.

Inclusion of DistR estimates into phylogenetic tree search of PHYML.—The experimental results when DistR estimates are incorporated into the ML tree search demonstrate the importance of accounting for different evolutionary pressures in phylogenetic inference.

Bootstrap support values for the ML tree using concatenated data are presented in Figure 7a. The bootstrap support for some of the clades was quite weak. Incorporating DistR estimates based upon both patristic and pairwise ML distances into the tree search led to the same ML tree, presented in 7b. Overall, bootstrap support was improved in most clades when DistR estimates were incorporated into the tree search.

The topology of the ML concatenation-based tree does not separate Zygomycota and Ascomycota as distinct clades, which is not surprising because the Zygomycota are traditionally difficult to place. Furthermore, the outgroup is incorrect since it should also contain *Homo sapiens* (which groups incorrectly with the zygomycete *Smittium* and the Ascomycota). This long-branch-attraction problem is due to the highly derived *Smittium* and *Homo* sequences. Using DistR estimates improves the bootstrap support in certain clades, and corrects the most evident topological problems, notably that Zygomycota more accurately group together (although as an unresolved paraphyletic group). Indeed, almost every branch that does not show 100% bootstrap support with the concatenated data have improved support when using protein rates. The only branching where support somewhat lessened from the concatenated to the protein-rate-based trees (and with using individual bootstrap rates) was the branching of *Allomyces* (a species that is difficult to place whatever the method or data set) with the remainder of the Chytridiomycota (Figs. 7a and b). Bootstrap support is strongest when using protein rates based upon pairwise ML distances, where the rate estimates were recomputed for each bootstrap replicate. This is perhaps because the variation in the pairwise ML distance rate estimates was smaller than, or on the same order of magnitude as, the rate estimates based on patristic ML distances.

Both the Kishino-Hasegawa (KH) test and Akaike Information Criterion (AIC) support the ML topology with protein rates as a better fit for the model to the data than the concatenated topology. Under the KH test (Kishino and Hasegawa, 1989, Shimodaira and Hasegawa, 2001); the concatenated topology was significantly worse than the DistR topology ($P < 0.0001$) when the topology was computed with rate estimates calculated based on both patristic and pairwise ML distances. The AIC provides a statistical measurement of the significance of the change in log-likelihood when using two different models to fit the data. The measure compensates for the increase in the number of parameters in the rates model. When DistR estimates based on pairwise ML distances are used, the AIC is 1043.65182 greater than the AIC for a single rate, concatenated analysis. When patristic ML distances are used for rate estimation, the increase in AIC over the concatenated analysis is 1068.7542. Both increases in AIC are very substantial, indicating that important information in the data that is disregarded by traditional concatenated analysis is captured by modeling protein rates.

CONCLUSION

A fast and accurate method to calculate the rates of partitioned data sets is presented. Although the analyses performed here are based upon protein sequence data, using nucleotide sequences should prove as effective. The error in the method is largely due to incorrect initial distance estimates for the proteins, which tend to be worse with smaller or poorly conserved sequences. Using pairwise ML distances for DistR estimation is just as accurate as using patristic ML distances. The estimates are accurate when compared to ML estimates and Bayesian posterior credible intervals for the rates. Incorporating the DistR estimates into PHYML leads to statistically better likelihood and topology.

ACKNOWLEDEGEMENTS

We thank Scott Bunnell, Alain Vandal, Tad Pupko, Tim Collins, and Olivier Gascuel for helpful comments on the manuscript. Thanks to Stéphane Guindon for kindly providing the source code of PHYML v2.2 for our use. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL), National Science and Engineering Research Council (NSERC grant 238975-01; DB), Fonds de recherche sur la nature et les technologies (FQRNT grant 2003-NC-81840; DB), and supply of laboratory equipment and informatics infrastructure by Genome Canada are gratefully acknowledged. RBB is supported by an NSERC PGS-B scholarship.

REFERENCES

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Nat. Acad. Sci. 99:1414–1419.

Bull, J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Bio. 42:384–397.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res. 31:1614–1623.

Bulmer, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8:868–883.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Cranston, K., and B. Rannala. 2005. Closing the gap between rocks and clocks. Heredity 94:461–462.

Eulenstein, O., D. Chen, J. G. Burleigh, D. Fernández-Baca, and M. J. Sanderson. 2004. Performance of flip supertree construction with a heuristic algorithm. Syst. Biol. 53:299–308.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylgenies. J. Mol. Evol. 53:447–455.

Felsenstein, J. 2004a. Inferring phylogenies, pages 148–149. Sinauer Associates, Sunderland, Massachusetts.

Felsenstein, J. 2004b. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle. URL: http://evolution.genetics.washington.edu/phylip.html

Gill, P., W. Murray, and M. Wright. 1982. Practical optimization. Academic Press.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: An application for the monte carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:559–560.

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

Huelsenbeck, J. P., J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Tree 11:152–158.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the braching order in hominoidea. J. Mol. Evol. 29:170–179.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr. Biol. 12:1773–1778.

Lapointe, F., and G. Cucumel. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306–312.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods: Empirical Bayesian methods are superior. Mol. Biol. Evol. 21:1781–1791.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Olsen, G. J., S. Pracht, and R. Overbeek. 1993. DNArates. URL: http://geta.life.uiuc.edu/gary/programs/DNArates.html.

Pupko, T., R. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002a. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77.

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002b. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Sumida, M., Y. Kanamori, H. Kaneda, Y. Kato, M. Nishioka, M. Hasegawa, and H. Yonekawa. 2001. Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the japanese pond frog *Rana nigromaculata*. Genes Genet. Systems 76:311–325.

Tanaka, J. S., and G. J. Huba. 1985. A fit index for covariance structure models under arbitrary GLS estimation. Br. J. Math. Statist. Psych. 38:197–201.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Res. 22:4673–4680.

Tomita, K., S. Yokobori, T. Oshima, T. Ueda, and K. Watanabe. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: Multiplied noncoding regions and transposition of tRNA genes. J. Mol. Evol. 54:486–500.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

First submitted 24 November 2004; reviews returned 18 March 2005; final acceptance 24 May 2005
Associate Editor: Tim Collins

APPENDIX 1

FORMULA FOR MEAN SQUARED ERROR AND GOODNESS-OF-FIT

Mean squared error is used to describe the accuracy of rate estimates. Because only relative rates can be computed rates are normalized so that the average rate over all proteins is one. Let $\tilde{r}$ denote the true rate (that is, the rate used in simulations), and let $\hat{r_1}$, ..., $\hat{r_{10}}$ be the rates estimated in the 10 replicates of the experiment. The mean squared error (MSE) is defined as

$$\frac{1}{10}\sum_{i=1}^{10}(\tilde{r}-\hat{r_i})^2.$$

Goodness-of-fit is used to measure the fit of the distance estimates to the distances in the tree used for simulation. There is a slight problem with scales since Pseq-Gen treats branch lengths as the expected number of substitutions per 100 sites while PHYML treats branch lengths as the expected number of substitutions per site. Let $\tilde{d}_{xy}^{(k)}$ be the distance between x and y in the tree used to simulate protein k, let r_k denote the rate used when simulating protein k, and let $\hat{d}_{xy}^{(k)}$ be the distance estimated by PHYML.

Given the differences in scale the goodness-of-fit measure used was

$$1.0-\frac{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)}-100\hat{d}_{xy}^{(k)}\right)^2}{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)}\right)^2}.$$

Note that the goodness-of-fit is at most one, and equals one if and only if there is a perfect fit.

APPENDIX 2

FAST ALGORITHM FOR LEAST-SQUARES ESTIMATION

This appendix shows how to quickly determine the vectors $\mathbf{p}$ and $\mathbf{r}$ that minimize the function $q(\mathbf{p}, \mathbf{r})$ in Equation (3)

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k} \right)^2$$

subject to the constraint that $h(\mathbf{p}) = \kappa$, where

$$h(\mathbf{p}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} p_{xy}$$

and κ is an arbitrary, positive constant. In the implementation of DistR

$$\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$$

which corresponds to the assumption that the unknown consensus distances are roughly centered on the average of the observed distances. This value can be computed in $O(nm^2)$ time for n proteins and m taxa. Any other positive constant will work, as the only effect is to change the scale of the rate estimates.

To simplify the mathematics substitute $s_k = \frac{1}{r_k}$ for each $k = 1, \ldots, n$. Let $\mathbf{s}$ denote the vector $[s_1, \ldots, s_n]^T$. Minimizing $q(\mathbf{p}, \mathbf{r})$ is then equivalent to minimizing

$$f(\mathbf{p}, \mathbf{s}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - s_k d_{xy}^{(k)} \right)^2. \tag{5}$$

Recall from calculus that the minimum of a one dimensional function can be found by determining where the first derivative is equal to zero. This condition extends to multidimensional functions with constraints. Refer to Gill et al. (1982) for an excellent introduction to the optimization tools used here.

If $(\mathbf{p}, \mathbf{s})$ together minimize the function f, subject to the condition that $h(\mathbf{p}) = \kappa$, then there exists a real number λ such that

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial p_{xy}} - \lambda \frac{\partial h(\mathbf{p})}{\partial p_{xy}} = 0 \quad & \text{for all taxa } x, y \\ \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial s_k} = 0 \quad & \text{for all proteins } k \\ h(\mathbf{p}) = \kappa. & \end{aligned} \tag{6}$$

In general, (6) is only a necessary condition for reaching the minimum, and not a sufficient condition. However, in this case the matrix formed from the second derivatives of $f(\mathbf{p}, \mathbf{s})$ is *positive definite*, so that the function f is convex (Gill et al., 1982). It follows that if $(\mathbf{p}, \mathbf{s})$ and λ satisfy (6) then $(\mathbf{p}, \mathbf{s})$ gives the global minimum.

It is possible to derive the partial derivatives of the functions f and h explicitly. To help with notation define the quantities:

$$\alpha_k = \sum_{xy} 2 w_{xy}^{(k)} \left(d_{xy}^{(k)} \right)^2 \quad \text{for all proteins } k;$$

$$\beta_{xy} = 2 \sum_{k=1}^{n} w_{xy}^{(k)} \quad \text{for all taxa } x, y;$$

$$\beta_{xy,k} = -2 w_{xy}^{(k)} d_{xy}^{(k)} \quad \text{for all proteins } k \text{ and taxa } x, y.$$

The partial derivative of f with respect to s_k, for some protein k, is

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial s_k} &= \sum_{xy} -2 w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k \right) d_{xy}^{(k)} \\ &= \alpha_k s_k + \sum_{xy} \beta_{xy,k} p_{xy}. \end{aligned}$$

The partial derivatives of f and h with respect to p_{xy}, for some taxa x, y, are

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial p_{xy}} &= \sum_{k=1}^{n} 2 w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k \right) \\ &= \sum_{k=1}^{n} \beta_{xy,k} s_k + \beta_{xy} p_{xy} \\ \frac{\partial h(\mathbf{p})}{\partial p_{xy}} &= \sum_{k=1}^{n} w_{xy}^{(k)} \\ &= \beta_{xy}/2. \end{aligned}$$

Note from the partial derivatives that the conditions in Equation (6) are linear equations involving the entries of $\mathbf{p}$, $\mathbf{s}$, and λ. As such, the next step is to rewrite 6 in terms of matrix algebra. Given that there are n proteins and m taxa define the following: let D be the $n \times n$ matrix with $\alpha_1, \alpha_2, \ldots, \alpha_n$ down the diagonal and zeros off the diagonal; let C be the $\frac{m(m-1)}{2} \times \frac{m(m-1)}{2}$ matrix with $\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}$ down the diagonal and zeros off the diagonal; let B be the $\frac{m(m-1)}{2} \times n$ matrix with rows indexed by unique pairs of taxa, columns indexed by proteins, and the entry corresponding to row xy and column k equal to $\beta_{xy,k}$; let $\mathbf{v}$ be the $\frac{m(m-1)}{2}$ dimensional vector $\mathbf{v} = \frac{1}{2}[\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}]^T$.

The conditions in Equation (6) can now be rewritten as

$$D\mathbf{s} + B^T \mathbf{p} = 0 \tag{7}$$

$$B\mathbf{s} + C\mathbf{p} + \mathbf{v}\lambda = 0 \tag{8}$$

$$\mathbf{v}^T \mathbf{p} = \kappa. \tag{9}$$

Define

$$\mathbf{u} = B^T C^{-1} \mathbf{v}$$
$$\omega = \mathbf{v}^T C^{-1} \mathbf{v}.$$

Solving for $\mathbf{p}$ in (8) gives:

$$\mathbf{p} = C^{-1}(-B\mathbf{s} - \mathbf{v}\lambda) \tag{10}$$

Substituting this into (9) and solving for λ gives:

$$\lambda = \frac{\kappa + \mathbf{v}^T C^{-1} B\mathbf{s}}{-\mathbf{v}^T C^{-1}\mathbf{v}}.$$
$$= \frac{\kappa + \mathbf{u}^T \mathbf{s}}{-\omega}.$$

Replacing λ with the above equation in (10) provides a solution for $\mathbf{p}$ in terms of the above defined matrices, vectors and $\mathbf{s}$ (i.e., there are no longer any unknowns except for $\mathbf{p}$ and $\mathbf{s}$):

$$\mathbf{p} = C^{-1}\left(- B\mathbf{s} + \mathbf{v}\frac{\kappa + \mathbf{u}^T\mathbf{s}}{\omega}\right) \tag{11}$$
$$= C^{-1}\left(\frac{\mathbf{v}\mathbf{u}^T}{\omega} - B\right)\mathbf{s} + \frac{\kappa}{\omega} C^{-1}\mathbf{v}. \tag{12}$$

Finally, substitute (12) into (7) to get

$$0 = D\mathbf{s} + B^T\mathbf{p}$$
$$= \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^T C^{-1} B\right)\mathbf{s} + \frac{\kappa}{\omega}\mathbf{u}.$$

Let

$$M = \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^T C^{-1} B\right).$$

Then, $\mathbf{s}$ is found by solving the equation:

$$M\mathbf{s} = -\frac{\kappa}{\omega}\mathbf{u}. \tag{13}$$

Consensus distances $\mathbf{p}$ are obtained by substituting $\mathbf{s}$ into Equation (12).

The entire computation is summarized in Appendix 3. The running time of the algorithm is $O(nm^2 + n^3)$ which is time optimal. The algorithm uses $O(n^2 + m^2)$ memory in addition to the $O(nm^2)$ required to store the distance estimates $d_{xy}^{(k)}$.

There are two complications that can arise in the above calculations. Firstly, it could be the case that for a particular pair of taxa x, y there is no single protein that contains both x and y. This means that β_{xy} is undefined, so that C is no longer invertible. This problem is easily solved. If there is no protein with both x and y then the line in (6) involving the partial derivative with respect to p_{xy} is satisfied trivially. Therefore, the row and column of C, the row of B, and entry of $\mathbf{v}$ indexed by the pair x, y can be removed. The reduced problem can be solved as before, although no estimate for p_{xy} is obtained. Row removal is handled in the pseudocode for the algorithm given in Appendix 3 by using constraints in the summations.

The second complication is that the optimization problem might have more than one solution, in which case the matrix M in (13) will not be invertible. This indicates that more information is required to estimate the relative rates, as would arise, for example, in a concatenation of two protein alignments over entirely different sets of taxa.

APPENDIX 3
THE DISTR ALGORITHM

Algorithm DISTR(d, w)
Input: Distance estimates $d_{xy}^{(k)}$ for each pair of taxa and each protein k.
Weights $w_{xy}^{(k)}$ for each distance estimate.
Missing distances have weight zero.
Output: Rate estimates r. Consensus distances p.

$\kappa = \sum_{k=1}^{n} \sum_{xy} w_{xy}^{(k)} d_{xy}^{(k)}$
for k from 1 to n do
  $\alpha_k \leftarrow \sum_{xy} 2 w_{xy}^{(k)} (d_{xy}^{(k)})^2$
  for all taxa x, y do
    $\alpha_{k,xy} \leftarrow -2 w_{xy}^{(k)} d_{xy}^{(k)}$
    $\beta_{xy,k} \leftarrow -2 w_{xy}^{(k)} d_{xy}^{(k)}$
for all taxa x, y do
  $\beta_{xy} \leftarrow 2 \sum_{k=1}^{n} w_{xy}^{(k)}$
$\omega \leftarrow \frac{1}{4} \sum_{xy} \beta_{xy}$
for k from 1 to n do
  $\mathrm{u}_k \leftarrow \sum_{xy} \beta_{xy,k}$
for k from 1 to n do
  $\mathrm{z}_k \leftarrow -\frac{\kappa}{\omega}\mathrm{u}_k$.
  for l from 1 to n do
    $M_{kl} \leftarrow -\sum_{xy:\beta_{xy}\neq 0} \frac{\beta_{xy,k}\beta_{xy,l}}{\beta_{xy}} + \frac{1}{\omega}\mathrm{u}_k\mathrm{u}_l$
    if $k = l$ then $M_{kl} \leftarrow M_{kl} + \alpha_k$
if M is nonsingular then output "Insufficient data to estimate rates"
solve $M\mathrm{s} = -\frac{\kappa}{\omega}\mathrm{u}$ to obtain s
for all taxa x, y such that $\beta_{xy} \neq 0$ do
  $\mathrm{p}_{xy} \leftarrow \sum_k \left(\frac{\mathrm{u}_k}{2\omega} - \frac{\beta_{xy,k}}{\beta_{xy}}\right)\mathrm{s}_k + \frac{\kappa}{2\omega}$
for k from 1 to n do
  $r_k \leftarrow \frac{1}{s_k}$
output r and p.