Estimating Phylogeny When Alignment is Uncertain*

NICK GOLDMAN[1]

Isaac Newton Institute for the Mathematical Sciences, University of Cambridge, Cambridge, UK

[1]Present address and address for correspondence:

Department of Genetics

University of Cambridge

Downing Street

Cambridge CB2 3EH

UK

tel: + 44 - (0)1223 - 333981

fax: + 44 - (0)1223 - 333992

e-mail: N.Goldman@gen.cam.ac.uk

Running head: GOLDMAN—PHYLOGENY WITH ALIGNMENT UNCERTAINTY

Keywords: [Alignment; maximum likelihood estimation; pairwise distances; phylogenetic estimation.]

*version 1.0; January 27, 1999; file v10.tex

Abstract.—Current methods for phylogenetic estimation from aligned sequences assume that the alignment is entirely correct. Regions of an alignment that are judged not to meet this assumption are discarded prior to phylogenetic analysis. This is typically done 'by eye'; there are no recognized or statistically justified methods, and any claimed 'noise reduction' or 'improvment in signal to noise ratio' is unquantifiable. Recent advances in sequence alignment techniques are leading to alignments which have associated probabilities of accuracy of each alignment column. I show how this additional information can be used in maximum likelihood and distance-based phylogenetic analyses in a mathematically justified manner, allowing an objective measure of alignment uncertainty to be incorporated. [Alignment; maximum likelihood estimation; pairwise distances; phylogenetic estimation.]

Current methods for phylogenetic estimation from aligned DNA or amino acid sequences make the assumption that the alignment is 'correct', i.e., that each column of the alignment contains residues that are homologous and shares a common ancestral history with every other column of the alignment. Consequently, sequence alignment is a prerequisite for phylogenetic estimation, and typically regions of an alignment that are not deemed to meet these assumptions are discarded prior to the phylogenetic analysis. The justification is that noise is removed from the phylogenetic signal; while some signal may be lost also, it is hoped that the reduction in noise will more than compensate for this. This process of rejection of alignment regions is generally performed 'by eye'.

Important exceptions to this are methods which simultaneously align sequences and perform phylogenetic estimations, via a model which describes both nucleotide substitutions and processes of insertion and deletion. Examples are the methods of Thorne et al. (1991, 1992) and Mitchison and Durbin (1995). The distinction between these methods and methods which iterate between alignment and phylogenetic estimation steps (e.g., Hein, 1990) is important here, since in the latter each application of the phylogenetic estimation step still assumes that the result of the preceeding alignment step is correct. Unfortunately, the simultaneous methods are not yet practical for realistic data sets.

Sequence alignment methods are becoming increasingly probabilistic, being based on models of substitution, insertion and deletion instead of on empirical heuristics. In particular it is becoming possible to estimate the probability that an alignment column is correct. The intention of this paper is to describe a method of phylogenetic analysis that can use such probabilities in a well-justified manner to calculate appropriately the contribution each alignment column should make. Rather than the subjective decision that a column is correct (and may be used for phylogenetic estimation) or wrong (and discarded), probabilities of being correct can be incorporated for all alignment columns. This permits an appropriate assessment of noise, without the need to discard columns and thus phylogenetic information. The method is simple, is applicable to existing maximum likelihood (ML) and distance-based phylogenetic analyses, and adds little computational expense to these methods.

Materials and Methods

The proposed new method for incorporating alignment uncertainty into phylogenetic estimation is most easily described in the terms normally used for ML estimation of phylogeny from aligned DNA sequences. Here, every possible nucleotide pattern b (e.g., the five-sequence pattern AACCG, indicating that the first two sequences have nucleotide A observed at a site, the next two sequences have C at that site, and the fifth sequence has G at that site) has a probability p_b defined by a model of nucleotide substitution and a candidate phylogeny. Calculation of probabilities p_b assumes that the patterns b are from sequences that are correctly aligned. If there are n sequences, then there are 4^n possible patterns b. A method for calculating the p_b has been described by Felsenstein (1981). A data set is described by the patterns b_i which occur at its sites $i = 1, 2, \ldots, N$ where N is the total number of sites in the alignment. The likelihood of the candidate phylogeny is then given by

$$L = c \cdot \prod_{i=1}^{N} p_{b_i} \tag{1}$$

or

$$\log L = \log c + \sum_{i=1}^{N} \log p_{b_i}, \tag{2}$$

with c an unimportant constant. ML methods estimate the phylogeny (tree topology and branch lengths, plus any other free parameters such as parameters of the nucleotide substitution model) by finding that candidate phylogeny which maximises $\log L$.

Now, suppose that we no longer assume that each site of the alignment is necessarily correct. We allow the possibility of events E_i, that the alignment is in error at site i, and their complements $\mathrm{E}_i^{\mathrm{C}}$, that the alignment is correct at site i. The probability that we observe pattern b_i at site i is now

$$\begin{aligned} \Pr(b_i) &= \Pr(b_i, \mathrm{E}_i^{\mathrm{C}}) + \Pr(b_i, \mathrm{E}_i) && (3) \\ &= \Pr(b_i \mid \mathrm{E}_i^{\mathrm{C}}) \cdot \Pr(\mathrm{E}_i^{\mathrm{C}}) + \Pr(b_i \mid \mathrm{E}_i) \cdot \Pr(\mathrm{E}_i) && (4) \\ &= p_{b_i} \Pr(\mathrm{E}_i^{\mathrm{C}}) + (1 - \Pr(\mathrm{E}_i^{\mathrm{C}})) \cdot \Pr(b_i \mid \mathrm{E}_i). && (5) \end{aligned}$$

(The first equality is derived from the law of total probability; the second and third are simple consequences of the laws of probability.)

Probabilistic measures of alignment accuracy available from some alignment methods (see below for examples) can give estimates, denoted f_i, for the values $\Pr(\mathrm{E}_i^{\mathrm{C}})$ for sequence alignments. The notation $\Pr(b_i \mid \mathrm{E}_i)$ indicates that in the case of alignment error we might consider the probability of the observed data to be a function of sequence position and/or the observed data at this position. Writing $g(b_i)$ for estimates of $\Pr(b_i \mid \mathrm{E}_i)$ (see below), equation 5 gives

$$\Pr(b_i) = f_i p_{b_i} + (1 - f_i) g(b_i). \tag{6}$$

Intuitively, we can understand equation 6 as follows: with probability f_i the alignment is correct at position i and then the probability of observing b_i is p_{b_i}; with complementary probability $(1 - f_i)$ the alignment is in error at position i and some other probability $g(b_i)$ is associated with observed data b_i. The numerical value of f_i in a sense 'down-weights' the contribution of the probability p_{b_i}, with greater effect the more likely it is that there is an error in alignment column i.

It is necessary to ensure that the total probability of observing anything at site i is equal to 1. This is achieved as follows. Since

$$\begin{aligned}
\Pr(\text{observe anything at site } i) &= \sum_{b_i=1}^{4^n} \Pr(b_i) && (7)\\
&= \sum_{b_i=1}^{4^n} (f_i p_{b_i} + (1 - f_i) g(b_i)) && (8)\\
&= f_i \sum_{b_i=1}^{4^n} p_{b_i} + (1 - f_i) \sum_{b_i=1}^{4^n} g(b_i) && (9)\\
&= f_i + (1 - f_i) \sum_{b_i=1}^{4^n} g(b_i) && (10)
\end{aligned}$$

then, setting the right hand side of equation 10 equal to 1, we require

$$\sum_{b_i=1}^{4^n} g(b_i) = 1. \tag{11}$$

I propose two schemes satisfying equation 11 for assigning values to the $g(b_i)$. The first is appropriate for the case that the observed pattern b_i is assumed to contain no relevant information when the alignment site is wrong. In this case, equation 11 results in the constraint that $g(b_i) = 4^{-n}$.

The second scheme is suitable for a case in which I assume that $\Pr(b_i \mid \mathrm{E}_i)$ is independent of sequence position i, but that the observed pattern b_i contains some

relevant information even when there is an error in alignment column i. I imagine the situation that, if an alignment column is wrong, the bases observed in that column share no common evolutionary history but are each independent observations of one sequence site evolving on the underlying phylogeny. For example, if an alignment column i containing the pattern AACCG is incorrect, g(AACCG) would be the probability of observing nucleotide A in one sequence; independently (i.e., at a different site) observing nucleotide A in a second sequence; independently observing C in both a third and fourth sequence; and independently observing G in a fifth sequence. In other words, the bases observed in that column are effectively random draws from the equilibrium frequencies of the nucleotides under the nucleotide substitution model currently in use. If pattern b_i is composed of nucleotides $b_{1i}b_{2i}\ldots b_{ni}$ (e.g., if b_i is AACCG then $b_{1i} = b_{2i} =$ A, $b_{3i} = b_{4i} =$ C and $b_{5i} =$ G), and if the equilibrium nucleotide frequencies are π_{X} for nucleotide X = A, C, G, T, then the appropriate form becomes:

$$g(b_i) = \prod_{j=1}^{n} \pi_{b_{ji}}. \tag{12}$$

For example, $g(\mathrm{AACCG} \mid \mathrm{E}_i) = \pi_{\mathrm{A}}^2\pi_{\mathrm{C}}^2\pi_{\mathrm{G}}$. This scheme satisfies equation 11 and has been used for all the examples in this paper. Note that this approach is computationally equivalent to assuming that, if the alignment column is incorrect, the bases in that column are related by a tree with infinitely long branches.

More complex schemes for assigning values to the $g(b_i)$ can be imagined. Even if there are some errors in an alignment column, some significant parts may be correct. However, to accomodate all the possibilities of this sort would require consideration of all possible evolutionary histories (shared or otherwise) of all nucleotides in the relevant alignment columns, which does not appear feasible except perhaps by a simulation approach. This has not yet been investigated.

Assuming independence of the $\Pr(\mathrm{E}_i^{\mathrm{C}})$ or their estimates f_i (see below), we can now write the likelihood of a candidate phylogeny as

$$L = c \cdot \prod_{i=1}^{N} \left(f_i p_{b_i} + (1 - f_i) g(b_i)\right) \tag{13}$$

or

$$\log L = \log c + \sum_{i=1}^{N} \log \left(f_i p_{b_i} + (1 - f_i) g(b_i)\right) \tag{14}$$

(cf. equations 1 and 2, and using equation 6). Likelihood maximisation can now proceed as usual, using the values of f_i supplied with the aligned data b_i and calculating the $g(b_i)$ according to the chosen scheme as above. The p_{b_i} are calculated in the normal manner, according to the chosen model of nucleotide substitution. In the examples below, the JC69 (Jukes and Cantor, 1969) and HKY85 (Hasegawa et al., 1985) models have been used. There is no reason why the method could not be extended to more complex models of nucleotide substitution, for example models incorporating a Gamma distribution to describe rate heterogeneity across sites (Yang 1994) or to models of amino acid replacements (see, e.g., Liò and Goldman, 1998).

Likelihood maximization over all possible tree topologies can be difficult for data sets containing many sequences. Often in such cases, distance matrix methods are found to be useful. The above methods can be adapted to these approaches also. I have been unable to derive a closed-form formula for pairwise distances (analogous to the familiar $d_{ij} = -(3/4)\log(1 - 4n_{ij}/3n)$ for the JC69 model) incorporating the f_i to measure alignment uncertainty. It is, however, simple and fast to estimate pairwise distances as the ML branch lengths of the (trivial) trees relating the sequence pairs. This can be done using the f_i as described above, and the resulting pairwise distances then used with the preferred distance-based phylogenetic estimation method. An example using this procedure is given below.

It is also possible to adapt the new methods described here to alignment columns containing gaps. Such columns are often removed prior to phylogenetic analysis, but it is possible to treat the gapped residues as missing data. This option is available, for example, in the ML programs in the PHYLIP package (Felsenstein, 1995), and is equivalent to considering the non-gapped residues in a column as having evolved on that part of the phylogeny under consideration which remains after all branches leading to the sequences containing gaps in that column are removed (Joseph Felsenstein, pers. comm.). The same approach can be implemented for the methods described here, with the appropriate $g(b_i)$ now being the product of equilibrium base frequencies evaluated only for the non-gapped residues in pattern b_i. An equivalent calculation is to assign the value 1 to π_-, using the symbol '-' for 'gap', and use the product over all residues,

including gaps, in column i. In this case,

$$g(b_i) = \prod_{\substack{j=1 \\ b_{ji} \neq -}}^{n} \pi_{b_{ji}} = \prod_{j=1}^{n} \pi_{b_{ji}}. \quad (15)$$

For example, $g(\text{AAC-G}) = \pi_{\text{A}}^2 \pi_{\text{C}} \pi_{\text{G}}$. An example of this approach is given below.

Results and Discussion

ML Estimation of PEPCK Phylogeny

The first example illustrates the use of the new methods described above for ML phylogeny estimation. A set of 18 Lepidopteran phosphoenolpyruvate carboxykinase (PEPCK) DNA sequences was used, as studied by Friedlander et al. (1996) and Goldman et al. (1998). These sequences are available by anonymous ftp from ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ds24063.dat. A new multiple alignment was created using a novel hidden Markov model method (Holmes and Durbin, 1998; Ian Holmes, pers. comm.). This method computes the probability that each site of each sequence is placed in the correct column of the multiple alignment; multiplying these probabilities across all non-gapped sites for each alignment column i gives the estimates f_i that the alignment column is correct.

The initial multiple alignment of the 18 PEPCK sequences, referred to as the PEPCK_{625} data set, consisted of 625 alignment columns. A second data set, PEPCK_{495}, was formed by removing all columns containing any gaps; this left 495 alignment columns. Graphical representations of the f_i for both PEPCK data sets are shown in Figure 1.

The two PEPCK data sets were analysed using the HKY85 model of nucleotide substitution. The phylogeny and the transition/ transversion rate ratio (κ) were estimated simultaneously by ML. Both data sets were analysed by both the new methods described above, using the measure of alignment uncertainty given by the values f_i, and by 'traditional' ML methods (ignoring the f_i or, equivalently, taking all the f_i to equal 1).

The ML tree topology is the same in all four cases, but there are differences in the lengths of branches in the trees. The optimal tree can be represented as ((((((Tci:a,

Tba), (Ldi, Tpe)), (Ese, Dgr:*b*)), (Kgr, Sar)), (Eau, Mca)), ((Mze, Aqu):*c*, Hps:*d*), (Bni, ((Aap:*e*, Cfe:*f*), (Tpa, Dme:*g*):*h*))), where the sequence labels are as described by Friedlander et al. (1996) and *a*–*h* represent certain branch lengths referred to in Figure 2 (other branch lengths not shown). For both PEPCK data sets, branch lengths tend to be smaller when the f_i are used to describe alignment uncertainty. Plots of the branch lengths estimated with and without the f_i (Fig. 2) indicate that the branch lengths are reduced by approximately a factor of 0.8 when the f_i are incorporated in the analysis, for both data sets.

Intuition suggests that branch lengths will generally be shorter in analyses using the f_i as described here. The more variation amongst sequences that is observed in a region, the harder it is to align that region accurately. Consequently, poor regions of an alignment (low f_i) will tend to be associated with regions of most divergence. If no use is made of the f_i, these high levels of divergence will simply be taken as evidence of long branches. However, if the associated relatively low values of f_i are used to down-weight these regions, the high levels of divergence are in effect being partly explained by possible alignment error, and have less tendency to inflate branch length estimates. It would be possible for branch lengths to be increased, for example if values of f_i were low in alignment regions where some sequences were very similar.

An undesirable effect would be to down-weight sites which were correctly aligned but happened by chance to be highly divergent. It is to be hoped that advanced alignment methods will be able to make this distinction. For example, I note that positions 119 and 120 of the PEPCK_{625} data set exhibit patterns TTTTTTTTTTTTTTTTTT and GCTGTCTACTCTTAGGCC respectively, yet $f_{119} = f_{120} = 0.979$.

Notice that the optimal log-likelihood values are considerably higher when the f_i are used, increasing by over 60 units for both the PEPCK_{625} and PEPCK_{495} data sets (Table 1). This increase is gained without adding any parameters to the model of sequence evolution—the f_i are part of the data, and the $g(b_i)$ are fully determined by the model of nucleotide substitution. However, no statistical tests have yet been attempted to test whether this is a significant improvement for these data. Table 1 also indicates that the estimates of κ are higher using the new method described here, but

there is insufficient evidence so far to draw any general conclusions from this.

ML Estimation of 5S rRNA Phylogeny and Pairwise Distances

The second example illustrates ML phylogeny estimation and bootstrap analysis and pairwise distance-based phylogeny methods using the new methods described above. A set of six 5S ribosomal RNA (rRNA) sequences was selected, from a bacterium, two plants, a fungus, an amphibian and an insect (see Fig. 3 for details). These are amongst the members of an example data set distributed with Hein's TREEALIGN alignment and phylogeny software (Hein, 1990; software and data available by anonymous ftp from ftp://ftp.ebi.ac.uk/pub/software/unix/treealign.tar.Z). A multiple alignment was created using the method of Holmes and Durbin (1998; Ian Holmes, pers. comm.) with estimated values f_i calculated as before. The original 5S rRNA sequences are all approximately 120 base pairs (bp) long; the multiple alignment consisted of 122 bp.

This data set (including gapped sites) was analysed by ML using the JC69 model of nucleotide substitution, both using the f_i to incorporate alignment uncertainty and ignoring the f_i. The resulting phylogenies are shown in Figure 3a, b. Again, there is no difference in the ML tree topology, but branch lengths tend to be slightly smaller when the f_i are used.

Additional analyses were performed to demonstrate other possible uses of the alignment uncertainty measures f_i. Noting that each f_i value applies to site i, and given the assumed independence of the f_i, it is possible to include them in bootstrap analyses (Felsenstein, 1985) simply by associating the appropriate value f_i with the data b_i whenever site i is selected for inclusion in a bootstrap data set. Such an analysis was performed for 1000 bootstraps for these 5S rRNA sequences, both incorporating and ignoring the f_i, and the results are given in Figure 3a, b. Notice that in this example the confidence assigned to the (true) groupings (plant1, plant2) and (insect, amphibian) are slightly increased by the use of the f_i. This suggests that the new method may be correctly reducing branch length estimates by discounting noise introduced by alignment errors, and consequently increasing ability to distinguish the correct evolutionary relationships. Of course, this one small data set is only meant as an example, and cannot be used to claim general efficacy for the new method. Even in cases where the new methods worked well, bootstrap proportions might increase or

decrease, depending on whether traditional methods were underestimating or overestimating the true signal extracted from the phylogenetic information.

In addition, all 15 possible pairwise alignments were generated using a method based on a model incorporating both nucleotide substitution and insertion and deletion events (Thorne, 1991, 1992; Jeff Thorne, pers. comm.). This method permits the calculation of the probability that each inferred alignment column of an optimal pairwise alignment exists in the true alignment; these probabilities gives the required values f_i appropriate for each pairwise alignment.

In this example, all sites of each pairwise alignment that contained a gap were removed. The pairwise alignments then each consisted of approximately 115 bp. A graphical representation of the f_i for the bacterium–amphibian and bacterium–fungus alignments is shown in Figure 4. ML distances between each pair were computed under the JC69 model of nucleotide substitution, both using the alignment uncertainty as measured by the f_i as described above, and by the usual method that ignores the f_i. The resulting pairwise distances are compared in Table 2 and Figure 5. Trees were estimated from these pairwise distance matrices using the method of Fitch and Margoliash (1967), abbreviated to FM. The resulting trees are shown in Figure 3c, d. Notice in this example that the tree topologies estimated by ML methods and FM methods both with and without use of the f_i are slightly different (Fig. 3). Notice also that all three different topologies are wrong: the ML methods are unable to resolve the (fungus, insect, amphibian) grouping; the FM methods resolve the (plant1, plant2, bacterium) grouping incorrectly. These differences are probably due to the short sequence lengths in this example, which is provided solely for illustrative purposes.

In this example, the down-weighting of uncertain alignment columns has again caused distance estimates to be smaller (Table 2; Fig. 5). As before, the explanation presumably is that uncertain regions of the pairwise alignments tend to be associated with regions of high divergence; down-weighting according to alignment uncertainty will tend to reduce the effect of these regions.

Conclusions

Molecular sequence alignments are not always correct. Errors introduce phylogenetic 'noise' which may be important in subsequent phylogenetic analyses

(Morrison and Ellis, 1997; Goldman, 1998). Previously, any allowance for this was performed subjectively, 'by eye', prior to phylogenetic analysis, typically by discarding some sites and asssuming all those retained to be 'correct'. The new method described here uses easily-understood probabilities that alignment columns are correct to make objective and appropriate allowance for the alignment uncertainty, and can be applied to ML or distance-based phylogenetic methods.

The allowance for alignment uncertainty is made by what can be considered a form of weighting (compare equations 2 and 14). Weight is given to the pattern b_i observed at site i (and taken from the phylogenetically uninformative alternative represented by $g(b_i)$) according to the probability f_i of its being genuinely derived from the underlying phylogeny. Sitewise weighting has been implemented in parsimony analyses for some time (Swofford et al., 1996), but takes a different form. There, the weights w_i (for each site i) are in effect used to indicate site multiplicities (or relative multiplicities): a site with weight 2 is in effect treated as two sites each of weight 1. An equivalent use of such weights in the probabilistic framework used above would replace equations 2 and 14 with $\log L = \log c + \sum_{i=1}^{N} w_i \log p_{b_i}$. Other than as a shorthand for alignment columns that are identical, it is difficult to know what such weights w_i mean, or how they would be derived.

Ideally, estimation of the f_i should be independent from site to site (see above) and independent also of the subsequent phylogenetic analysis. In practice, the f_i are calculated from the data; it is hoped that the information the alignment methods use in doing this is largely independent of the information used in subsequent phylogenetic analyses. In this manner, the method begins to approach the simultaneous alignment and phylogenetic estimation methods of Thorne et al. (1991, 1992) and Mitchison and Durbin (1995). These use the information in a set of sequences more fully, via a unified model of nucleotide substitution and insertion/ deletion. In particular, they permit plausible alignments (or, rather, potential hypotheses of the relatedness of sites of each sequence) other than the single optimal one to make a contribution to the phylogenetic estimation. These methods remain computationally impractical for real data sets, however, and while this remains the case the methods described in this paper may find some use.

Data Availability

The PEPCK and 5S rRNA data sets studied, including the values of the f_i, are available from the author via http://ng-dec1.gen.cam.ac.uk/ftree/index.html.

Acknowledgments

The question I have addressed here was proposed to me by Temple Smith during the Isaac Newton Institute for the Mathematical Sciences programme *Biomolecular Function and Evolution in the Context of the Genome Project* (July–December 1998). I am grateful to Ian Holmes and Jeff Thorne for their considerable assistance with the sequence alignments used as examples, and to Ziheng Yang, Ian Holmes and Jeff Thorne for comments on an earlier version of this paper. This work was supported by a Visiting Fellowship at the Isaac Newton Institute for the Mathematical Sciences, Cambridge, awarded to the author and funded by EPSRC Grant GR K99015.

References

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

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

FELSENSTEIN, J. 1995. PHYLIP (Phylogenetic inference package), version 3.57. Univ. Washington, Seattle.

FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279–284.

FRIEDLANDER, T. P., J. C. REGIER, C. MITTER, AND D. L. WAGNER. 1996. A nuclear gene for higher-level phylogenetics—phosphoenolpyruvate carboxykinase tracks Mesozoic-age divergences within *Lepidoptera* (Insecta). Mol. Biol. Evol. 13:594–604.

GOLDMAN, N. 1998. Effects of sequence alignment procedures on estimates of phylogeny. BioEssays 20:287–290.

GOLDMAN, N., J. L. THORNE, AND D. T. JONES. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458.

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.

HEIN, J. J. 1990. A unified approach to alignment and phylogenies. Meth. Enz. 183:626–645.

HOLMES, I., AND R. DURBIN. 1998. Dynamic programming alignment accuracy. J. Comput. Biol. 5:493–504.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

LIÒ, P., AND N. GOLDMAN. 1998. Models of molecular evolution and phylogeny. Gen. Res. 8:1233–1244.

MITCHISON, G., AND R. DURBIN. 1995. Tree-based maximal likelihood substitution matrices and hidden Markov models. J. Mol. Evol. 41:1139–1151.

MORRISON, D. A., AND J. T. ELLIS. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. Mol. Biol. Evol. 14:428–441.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

THORNE, J. L., H. KISHINO, AND J. FELSENSTEIN. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114–124.

THORNE, J. L., H. KISHINO, AND J. FELSENSTEIN. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34:3–16.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

TABLE 1. PEPCK analyses. All analyses are by ML under the HKY85 model. The maximum value of the log-likelihood, $\log \hat{L}$, and the ML estimate of the transistion/transversion ratio, $\hat{\kappa}$, are reported.

dataset	analysis method	$\log \hat{L}$	$\hat{\kappa}$
PEPCK_{625}	f_i used	-7533.3	2.17
PEPCK_{625}	f_i ignored	-7595.0	1.96
PEPCK_{495}	f_i used	-6000.1	2.02
PEPCK_{495}	f_i ignored	-6064.3	1.85

TABLE 2. 5S rRNA pairwise distances. Sequence labels are as given in Fig. 3. Distances above the diagonal were calculated using the probabilities f_i; distances below the diagonal do not use the f_i. All are ML distances using the JC69 model, and are reported in terms of the expected number of substitutions per site.

	bacterium	plant1	plant2	fungus	insect	amphibian
bacterium	0	0.250^a	0.162^b	0.844^c	0.693	1.266^d
plant1	0.579^a	0	0.189^e	0.664	0.616	0.526
plant2	0.403^b	0.191^e	0	0.433^f	0.488	0.497
fungus	0.966^c	0.664	0.572^f	0	0.451	0.483
insect	0.750	0.616	0.516	0.451	0	0.358
amphibian	1.390^d	0.526	0.499	0.483	0.356	0

$^{a-f}$The pairs of distances indicated a–f are also indicated in Figure 5.

FIGURE LEGENDS

FIGURE 1. Probabilities f_i for the PEPCK_{625} and PEPCK_{495} data sets. Grey circles show the values of f_i for sites $i = 1, \ldots, 625$ of the PEPCK alignment. Black points indicate the 495 of these sites which are included in the PEPCK_{495} data set, i.e., which contain no gaps.

FIGURE 2. Comparisons of branch lengths in the PEPCK phylogenies. The y-axis gives the estimated branch length when the probabilities f_i are used to 'down-weight' sites according to the probability that they are incorrectly aligned, and the x-axis gives the corresponding branch length estimate when the f_i are not used. Points labelled a–h are as defined in the text. (a) Branch lengths of the phylogeny estimated from the PEPCK_{625} data set. (b) Branch lengths of the phylogeny estimated from the PEPCK_{495} data set.

FIGURE 3. Phylogeny estimates for the 5S rRNA data set. Sequence labels are as follows: bacterium—*Bacillus pasteurii*, plant1—*Equisetum arvense*, plant2—*Secale cereale*, fungus—*Auricularia auricula-judae*, insect—*Drosophila melanogaster*, amphibian—*Xenopus laevis*. Branch lengths are all drawn relative to the common scale bar, which indicates a distance of 0.2 nucleotide substitutions expected per site. (a) ML phylogeny using f_i. Bootstrap values of 89% and 100% were observed for two internal branches as indicated. The grouping (fungus, insect, amphibian) was unresolved in 26% of 1000 bootstrap replicates; the true resolution, (fungus, (insect, amphibian)), was recovered in 18% of replicates. (b) ML phylogeny, f_i not used. Bootstrap values are analogous to (a), with the true resolution (fungus, (insect, amphibian)) recovered in 15% of replicates. (c) FM phylogeny using f_i. The plant2 sequence lies directly on the branch indicated. (d) FM phylogeny, f_i not used.

FIGURE 4. Probabilities f_i for two pairwise alignments of 5S rRNA sequences. Grey circles show the values of f_i for sites $i = 1, \ldots, 117$ (after gapped sites were removed) of the fungus-bacterium alignment. Black points indicate the values for sites $i = 1, \ldots, 117$ (after gapped sites were removed) of the amphibian-bacterium alignment. (Note that this numbering scheme does not guarantee that a site in the fungus-bacterium alignment corresponds to the same-numbered site in the amphibian-bacterium alignment.)

FIGURE 5. Comparisons of pairwise distances for the 5S rRNA sequences. The y-axis gives the estimated pairwise distance when the probabilities f_i are used to 'down-weight' sites according to the probability that they are incorrectly aligned, and the x-axis gives the corresponding pairwise distance estimate when the f_i are not used. Points are shown for all 15 possible pairwise comparisons, including those labelled a–f in Table 2.

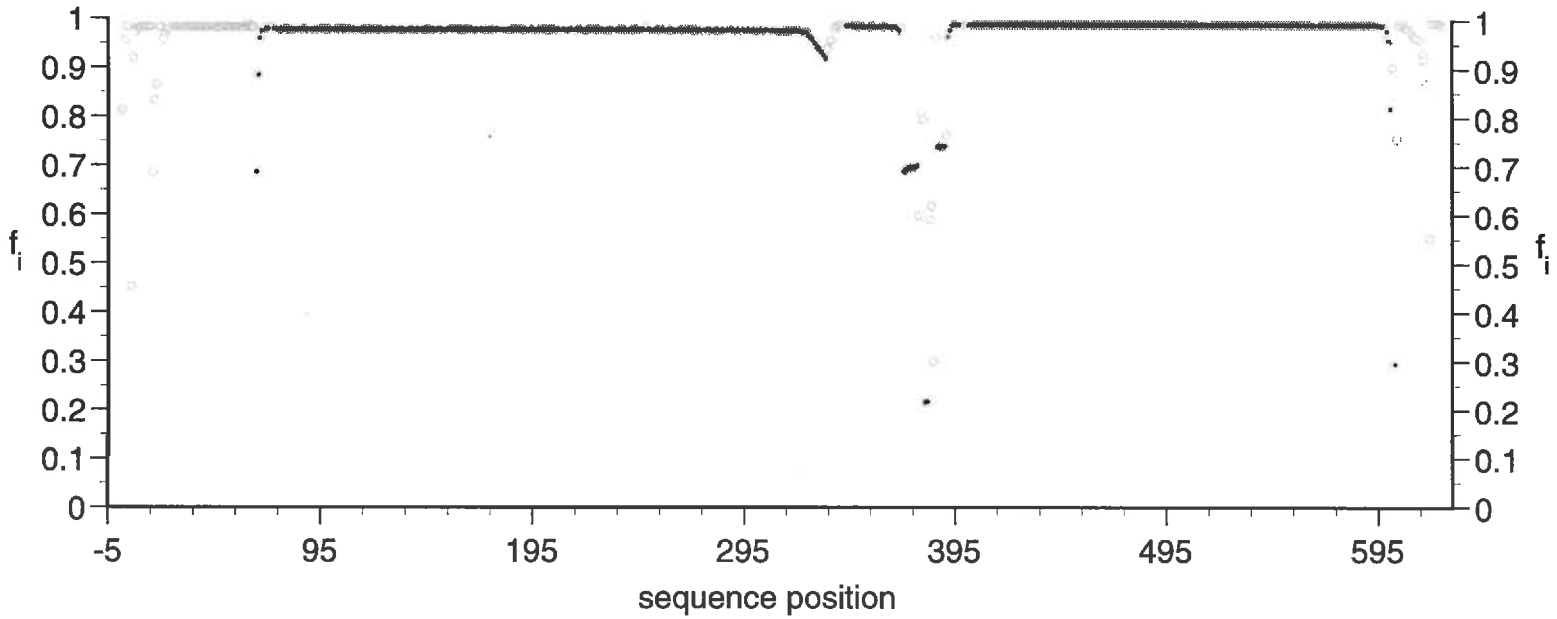
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
f_i
-5
95
195
295
395
495
595
sequence position

Fig. 1

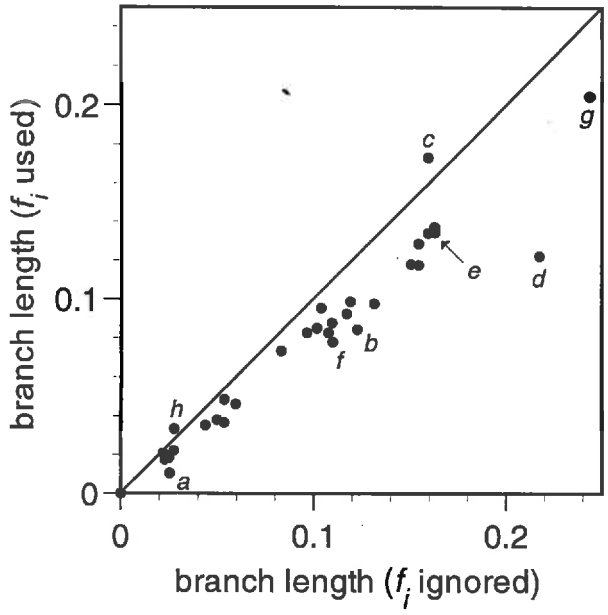
branch length (f_i used)
0.2
0.1
0
0
0.1
0.2
branch length (f_i ignored)
a
b
c
d
e
f
g
h

Fig (2a)

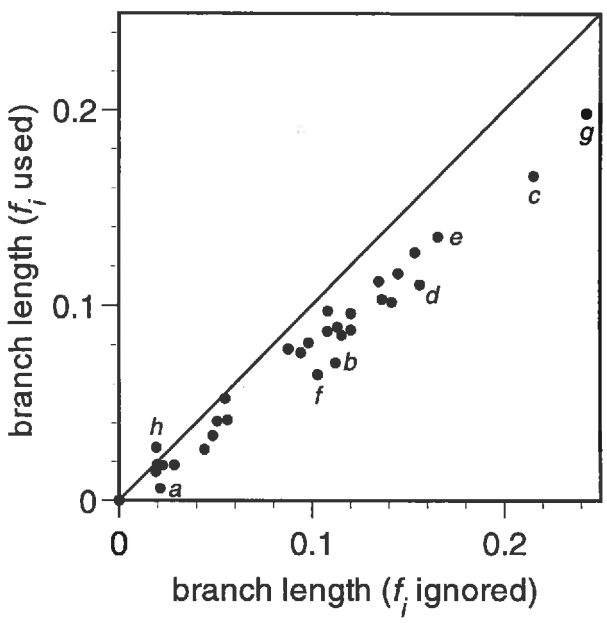
branch length (f_i used)
0.2
0.1
0
0
0.1
0.2
branch length (f_i ignored)
a
b
c
d
e
f
g
h

Fig. (2b)

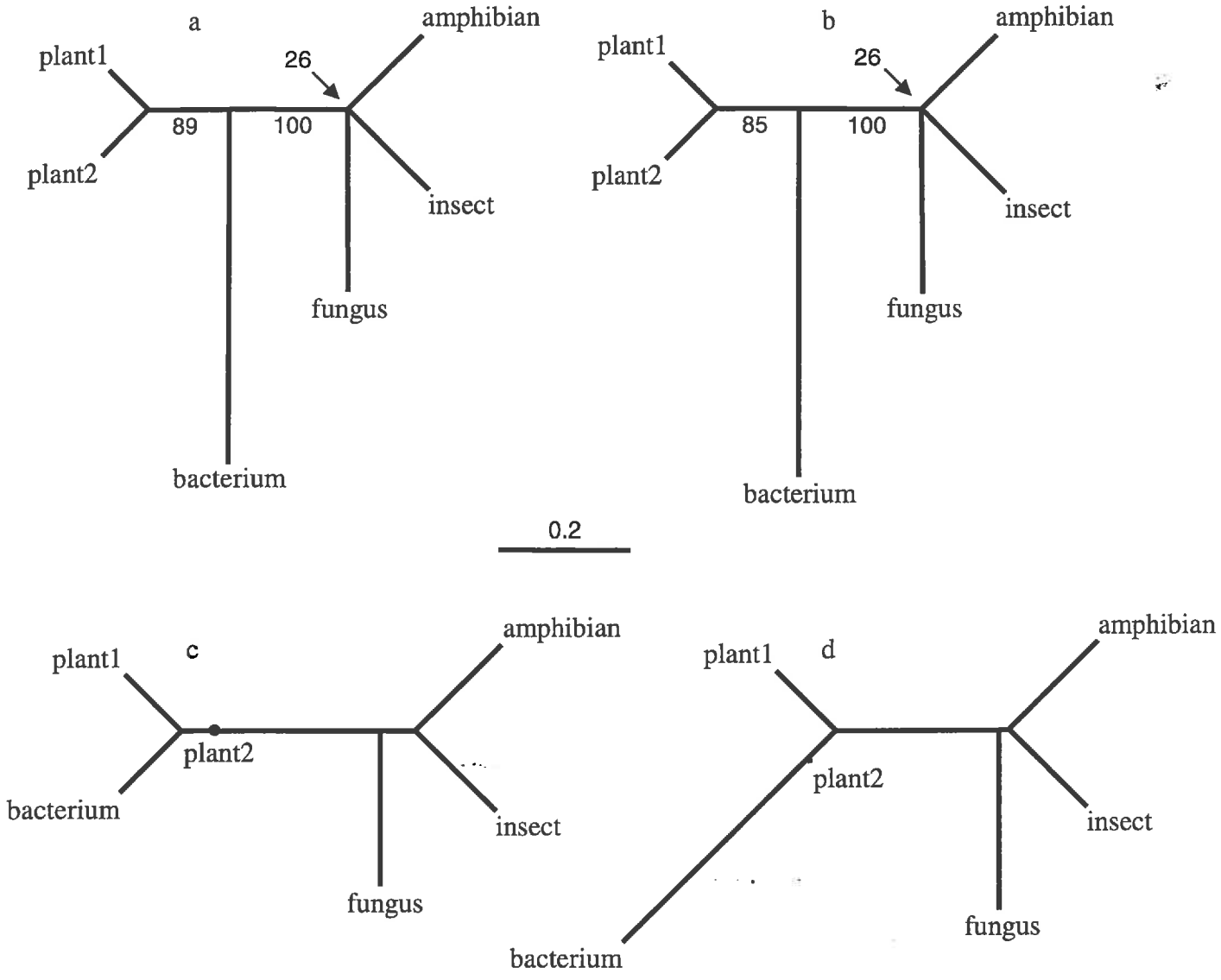
a
plant1
plant2
26
89
100
amphibian
insect
fungus
bacterium
b
plant1
plant2
26
85
100
amphibian
insect
fungus
bacterium
0.2
c
plant1
plant2
bacterium
amphibian
insect
fungus
d
plant1
plant2
amphibian
insect
fungus
bacterium

Fig (3)

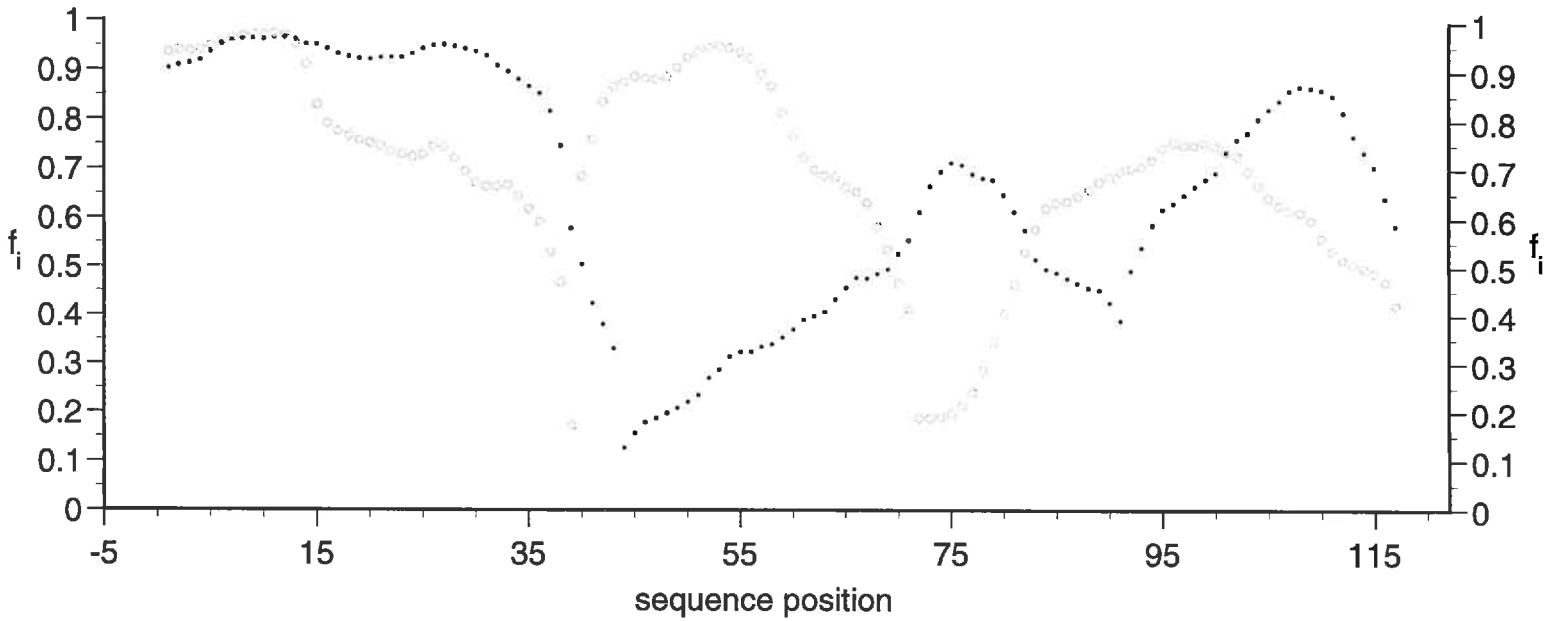
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
f_i
-5
15
35
55
75
95
115
sequence position

Fig. 4)

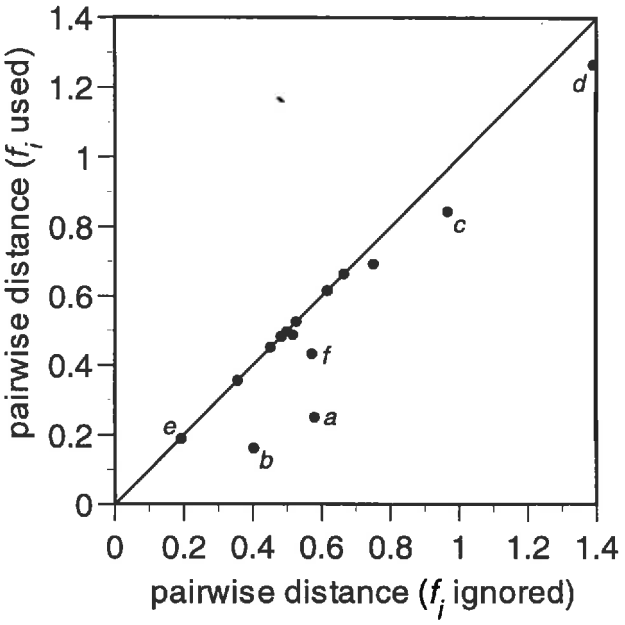
1.4
1.2
1
0.8
0.6
0.4
0.2
0
pairwise distance (f_i used)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
pairwise distance (f_i ignored)
a
b
c
d
e
f

Fig. (5)