Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution

Nick Goldman,* Jeffrey L. Thorne[†] and David T. Jones[‡]

*Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom, [†]Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 and [‡]Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom

Manuscript received October 5, 1997
Accepted for publication February 17, 1998

ABSTRACT

Empirically derived models of amino acid replacement are employed to study the association between various physical features of proteins and evolution. The strengths of these associations are statistically evaluated by applying the models of protein evolution to 11 diverse sets of protein sequences. Parametric bootstrap tests indicate that the solvent accessibility status of a site has a particularly strong association with the process of amino acid replacement that it experiences. Significant association between secondary structure environment and the amino acid replacement process is also observed. Careful description of the length distribution of secondary structure elements and of the organization of secondary structure and solvent accessibility along a protein did not always significantly improve the fit of the evolutionary models to the data sets that were analyzed. As indicated by the strength of the association of both solvent accessibility and secondary structure with amino acid replacement, the process of protein evolution—both above and below the species level—will not be well understood until the physical constraints that affect protein evolution are identified and characterized.

WIDELY used models of sequence evolution provide notoriously poor fits to actual sequence data (*e.g.*, Goldman 1993; Goldman and Yang 1994). For example, the infinite sites model of population genetics assumes that distinct mutational events never affect the same site in a DNA sequence. This assumption cannot be reconciled with the data: this model predicts that at most two nucleotide types will be represented in a column of a sequence alignment, yet it is often the case that an actual aligned data set contains one or more alignment columns where three or sometimes all four nucleotide types are represented. For these data sets, the infinite sites model can be rejected by inspection. More sophisticated models (*e.g.*, Jukes and Cantor 1969; Hasegawa *et al.* 1985; Yang 1994) are typically employed when sequences from different species are being compared. These may be superior to the infinite sites model in that they cannot be rejected by inspection, but these also tend to be rejected by goodness-of-fit tests such as the one proposed by Goldman (1993).

Although their assumptions may be violated and there may be statistical grounds for rejecting them, simple models of sequence evolution may still be adequate for investigating the questions to which they are often applied. In population genetics, the infinite sites model may be good enough for testing the neutral hypothesis of Kimura (1983). In macroevolutionary applications, the Jukes-Cantor (1969) model may often be sufficient for accurate reconstruction of phylogenies.

Nevertheless, the limitations of widely used models of sequence evolution often prevent more refined and informative questions from being addressed. A key to modelling and understanding the evolutionary process is identification and characterization of the constraints that evolution "perceives" as proteins diverge. Selective constraints on protein structure would be expected to give rise to associations between patterns of amino acid replacement and structure. In this article, we extend our earlier approach (Thorne *et al.* 1996) for characterizing these associations by increasing the number of secondary structures distinguished by the evolutionary model and by considering solvent accessibility (*i.e.*, whether or not a residue is near the surface of a protein and relatively exposed to solvent). We also add a more realistic description of secondary structure organization along a protein sequence.

There is a tradition of studies (*e.g.*, Overington *et al.* 1990; Lüthy *et al.* 1991; Topham *et al.* 1993; Wako and Blundell 1994) that attempts to associate amino acid replacement patterns and protein secondary structure or solvent accessibility, but these studies were performed without direct consideration of evolution and their findings are therefore difficult to apply to evolutionary questions. The work of Koshi and Goldstein (1995) is more interpretable from an evolutionary perspective but was not aimed at the study of evolution and instead addressed prediction of secondary structure and solvent

Corresponding author: Nick Goldman, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. E-mail: n.goldman@gen.cam.ac.uk

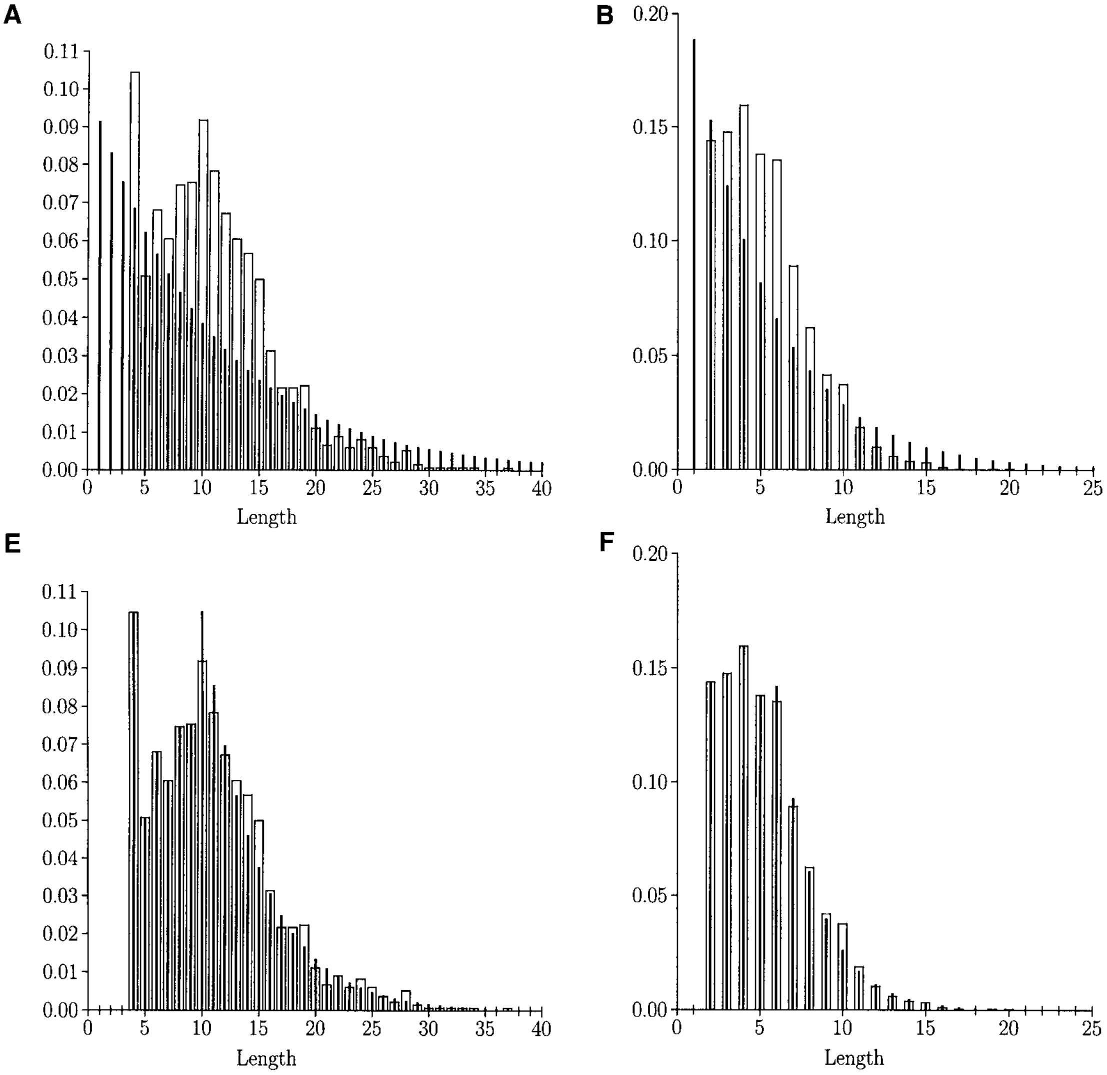
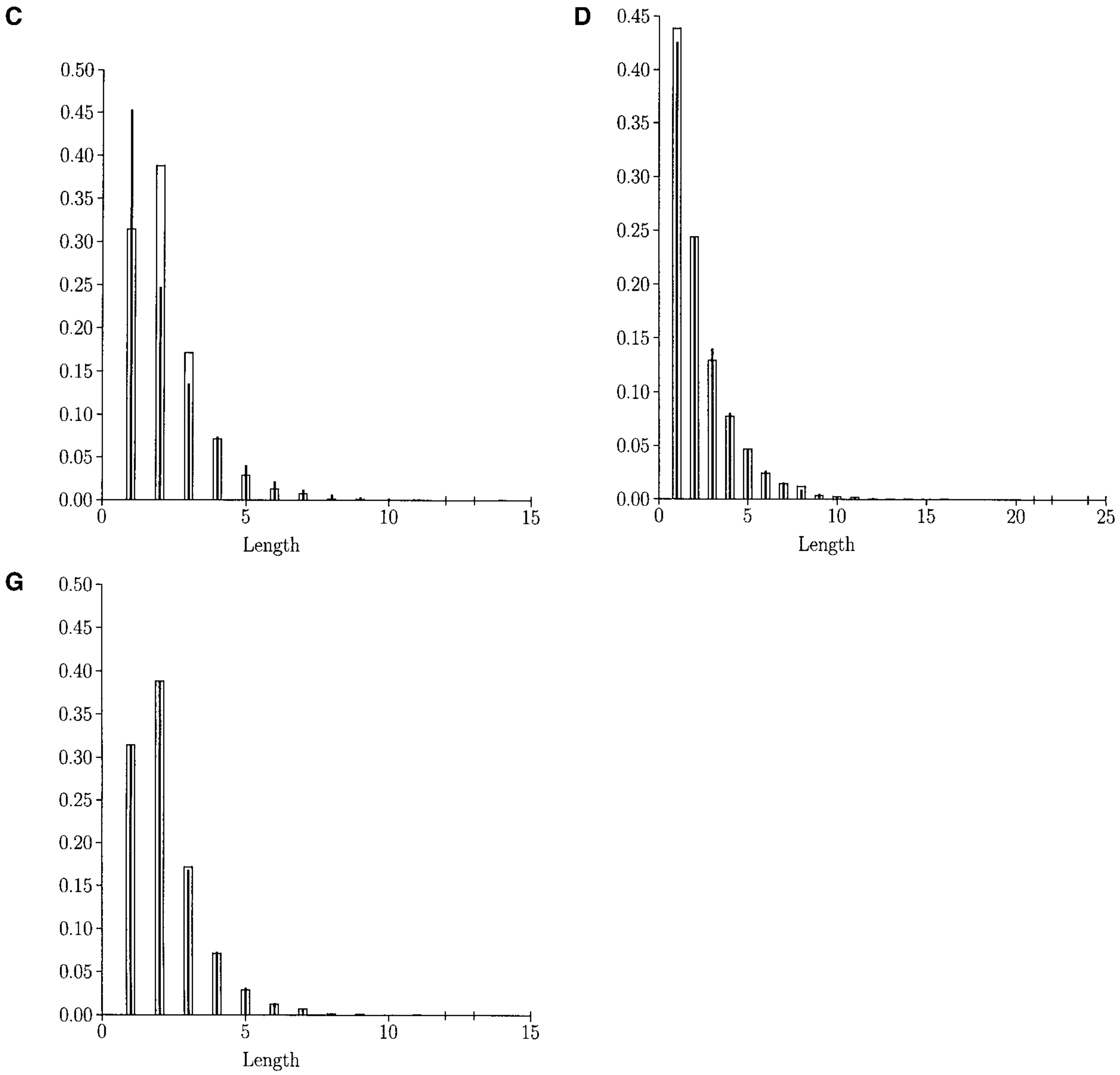


Figure 1.—Observed and predicted secondary structure length distributions. Open bars are empirically observed frequencies of secondary structure lengths. Solid bars are predicted frequencies. The predicted frequencies for Figures 1 (A–D) assume a geometric length distribution. The predicted frequencies for Figures 1 (E–G) are those used by the HMM. For coils, the HMM uses the geometric distribution depicted in (D). (A) Observed and geometric distributions of helix lengths (n = 1339). (B) Observed and geometric distributions of sheet lengths (n = 1862). (C) Observed and geometric distributions of turn lengths (n = 4482). (D) Observed and geometric distributions of coil lengths (n = 5113). (E) Observed and HMM distributions of helix lengths (n = 1339). (F) Observed and HMM distributions of sheet lengths (n = 1862). (G) Observed and HMM distributions of turn lengths (n = 4482).

accessibility. Other recently proposed models incorporate variation of preferred amino acid residues among sites (*e.g.*, Brown *et al.* 1993; Bruno 1996) but are not designed to facilitate study of the relationship between protein structure and evolution. With the approaches described here, associations between protein secondary structure, solvent accessibility, and the pattern of protein evolution can be investigated from within a likelihood inference framework.

We assume that each site in a protein belongs to one of several categories. The categories need to be predetermined, but the category to which each site belongs does not need to be prespecified. In this study, we have four types of secondary structure: α-helix, β-sheet, turn, and coil. For each secondary structure type, we further divide sites into those that are relatively exposed to solvent and those that are buried.

Having two accessibility classes, buried (b) and exposed (e), for each of the four secondary structure types (H, E, T, C) gives rise to a model with protein sites that belong to one of eight categories (Hb, He, Eb, Ee, Tb, Te, Cb, Ce). We use a hidden Markov model (HMM) approach (see Churchill 1989; Asai *et al.* 1993; White *et al.* 1994; Yang 1995; Felsenstein and Churchill 1996) to describe organization of these eight categories along a protein sequence, and we describe the length distribution of secondary structures more accurately than did our earlier model. The model that results from



these changes is both relatively complicated and comparatively realistic when contrasted with previous explicit models of protein sequence evolution. In the sections that follow, we first describe the model. We then investigate several data sets to understand how much various features of the model improve its fit.

MATERIALS AND METHODS

BRKALN database: Fixed parameters of the model were estimated as described below from a database of structure-related amino acid sequence alignments. The BRKALN database maintained by D.T. Jones (unpublished results) contains amino acid sequences classified into families of closely related sequences for which the tertiary structure of at least one member has been experimentally determined. The database is built by extracting nonhomologous sequences from the Brookhaven Protein Databank (PDB; Bernstein *et al.* 1977). Low resolution (>2.6Å) and NMR structures are excluded. Each sequence is compared to the OWL nonredundant protein sequence database (Bleasby and Wootton 1990) with a dynamic programming-based similarity search (Gotoh 1982) to find sequences of >30% identity with the sequence of known structure. Each family found in this manner is aligned via the multiple sequence alignment method of Taylor (1988).

Secondary structure assignments and solvent accessibility scores are calculated for the protein of known structure with the DSSP program (Kabsch and Sander 1983) and are extrapolated across each aligned sequence family by assuming that all residues in an alignment column share the secondary structure and solvent accessibility classification of the homologous residue in the protein with experimentally determined structure. Residues inserted into the middle of a secondary structure element are assigned that secondary structure, and insertions elsewhere are defined as having unknown secondary structure. Using the January 1995 release of PDB and release 25.0 of OWL resulted in the BRKALN database containing 207 families of sequences.

The DSSP assignments that occur in the BRKALN database are classified as follows: H, helix; E, sheet; S and T, turn; all others (*i.e.*, B, ".", G, and I), coil. The decision to classify S and T as a separate turn category instead of as coil (as in Thorne *et al.* 1996) was made because the DSSP assignments S (bend) and T (turn) yield amino acid replacement patterns

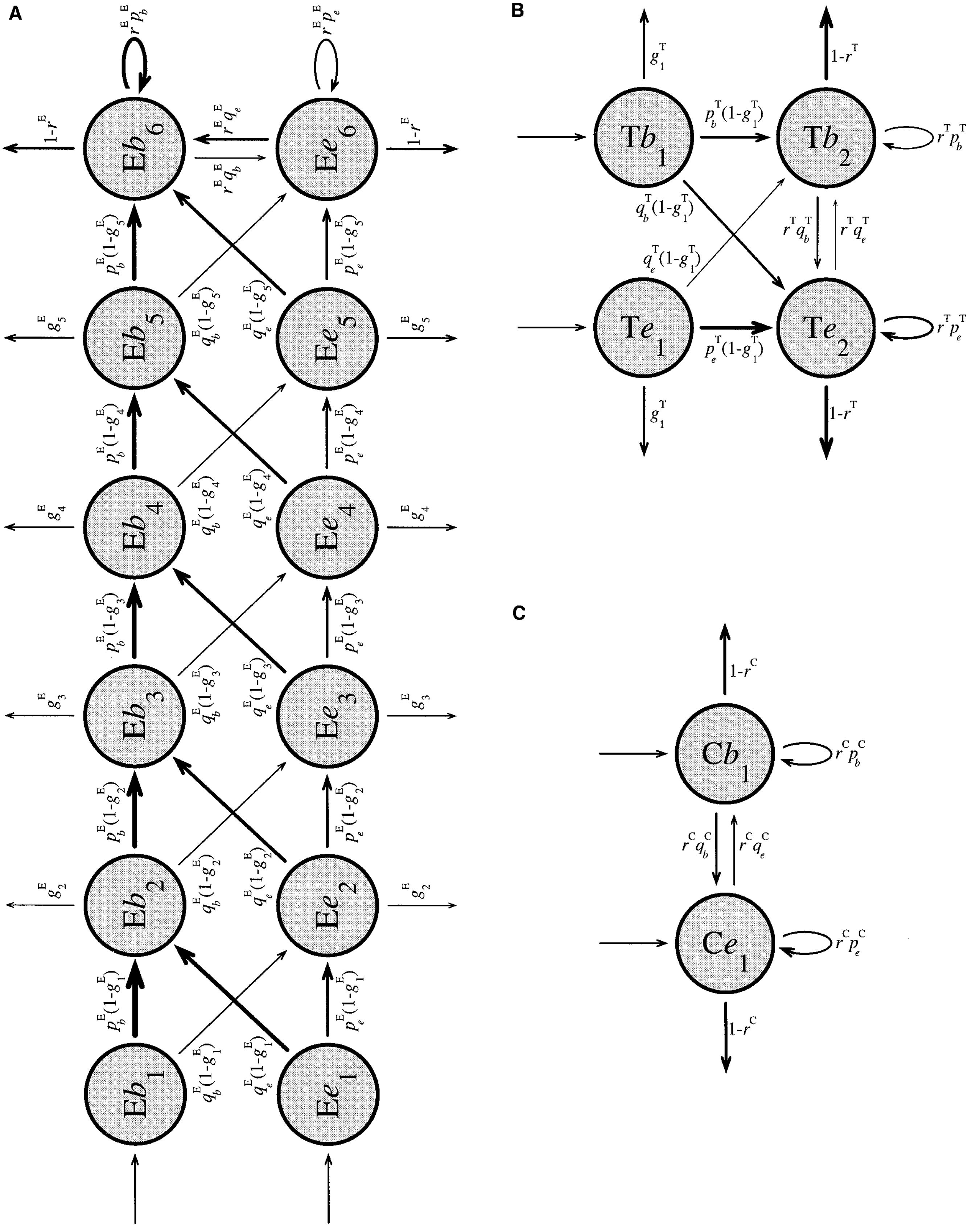


Figure 2.—Examples of permitted transitions among the 38 HMM states. Arrows indicate the permitted transitions among the states illustrated and are labelled with the parameter combinations that define the transition probabilities ρ_{ij}. Thicknesses of arrows are approximately proportional to the values of the corresponding ρ_{ij}. (A) Twelve sheet states. (B) Four turn states. (C) Two coil states. (D) All permitted transitions from the Eb_3 state. The HMM may progress to either of the next sheet states (Eb_4 or Ee_4) or may leave the sheet and enter a helix, turn, or coil via their first states (Hb_1 and He_1, Tb_1 and Te_1, or Cb_1 and Ce_1).

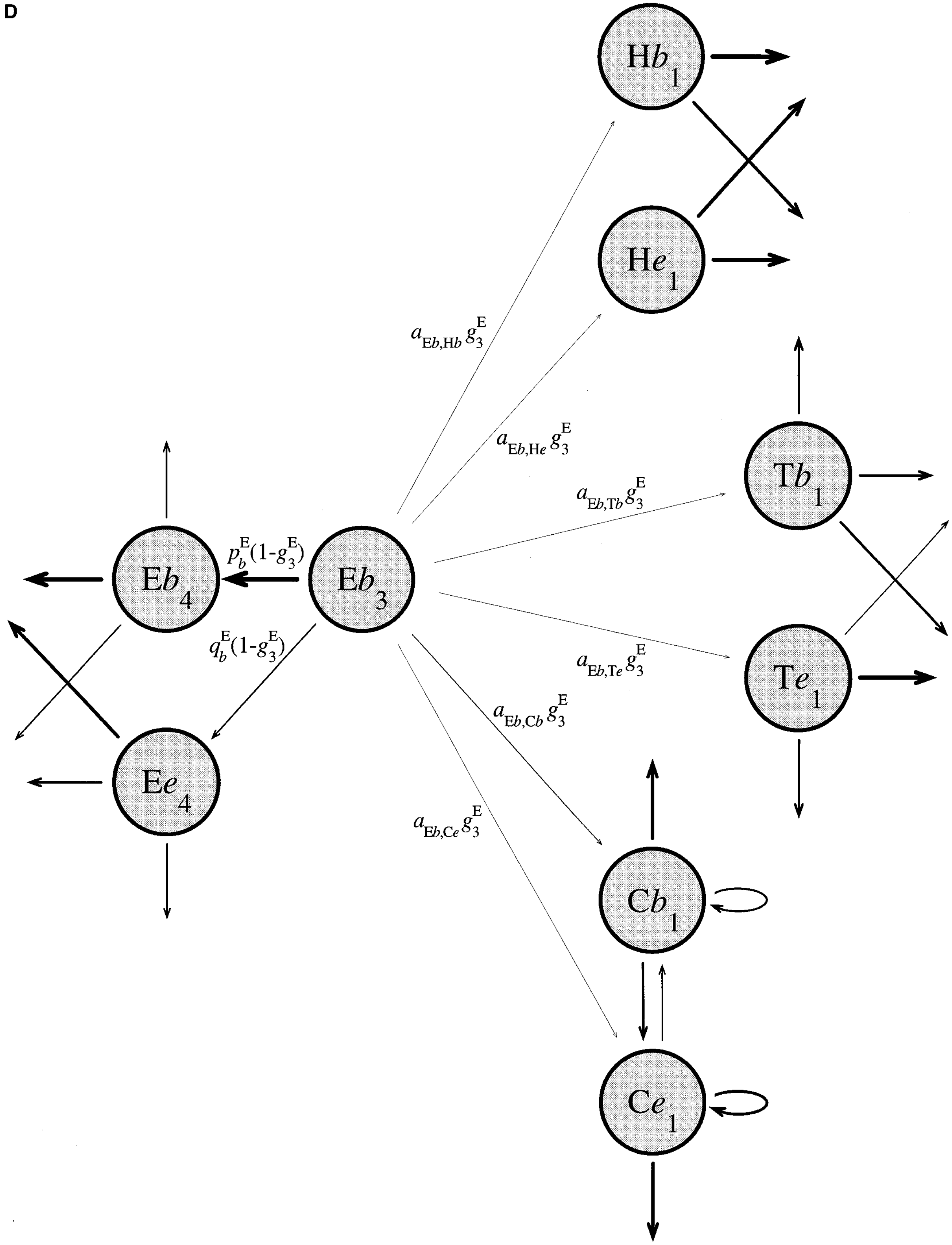


Figure 2.—*Continued*

and rates that are qualitatively similar to one another. DSSP solvent accessibility values were converted to two states, buried (accessibility <10%) and exposed (accessibility ≥10%). Relative accessibility was estimated by using a fully extended Gly-X-Gly tripeptide as the reference state. This resulted in approximately half of the sites in known structures of the BRKALN database being classified as buried and half as exposed. There is scope for more refined solvent accessibility schemes to be explored in the future.

Replacement processes: Each combination of secondary structure and accessibility status is associated with a particular amino acid replacement process. As is the case for most widely used models of nucleotide substitution and amino acid replacement, our models of amino acid replacement are Markovian with respect to time. To specify the process of evolution for a specific replacement category *k* that is one of the eight in our model, there are parameters α_{ij}^{k}, the relative rate of change from type *i* to *j*. Slight modifications (Jones *et al.* 1992; Goldman *et al.* 1996; Thorne *et al.* 1996) of the approach by Dayhoff *et al.* (1972, 1978) are used to base estimates of α_{ij}^{k} on observed amino acid replacement counts. These counts are made by comparing pairs of sequences in the BRKALN database that are 85% or more identical and by recording the number of times for replacement category *k* that amino acid type *i* is observed in one sequence of the pair and type *j* is observed at the corresponding site in the other.

Hidden Markov model: Neither the secondary structure nor the solvent accessibility at one site in a protein is independent of that at nearby sites. In fact, secondary structures at adjacent sites are strongly positively correlated. Accessibility status at adjacent sites is also positively correlated, although less strongly so. We adopt a hidden Markov model (HMM) to describe the organization of secondary structure and accessibility status along amino acid sequences. The states of the model correspond to the underlying but unobserved (hence hidden) secondary structure and accessibility. Transitions among the states are modelled with a Markov process.

A very simple HMM would ignore accessibility status and only describe the organization of the four secondary structure types along a protein sequence. A consequence would be that the number of sites in each uninterrupted stretch of a secondary structure would be geometrically distributed. Figure 1, A–D, compares the length distributions predicted under this simple model for the four secondary structure types with the empirically observed frequencies in the 207 known structures of the BRKALN database. For the secondary structure type coil (Figure 1D), the observed and predicted length distributions are reasonably close. For the other three secondary structure types, the observed distribution and that predicted under this simple HMM are quite different. For example, the geometric distribution necessarily has length one as its mode, but it is impossible for an α-helix to consist of just one residue.

To provide a better fit between observed and predicted distributions, our HMM has states for the first, second, *etc.*, positions in a secondary structure. We illustrate this with sheets. To represent the *i*th position of a sheet for $i \in \{1,2,...,5\}$, we use Eb_i or Ee_i, respectively, when the accessibility status is buried or exposed. The states Eb_6 and Ee_6 denote buried and exposed sites in position six or greater of a sheet.

We constrain the HMM to enter a sheet only through either the Eb_1 or Ee_1 states. Once in state Eb_i or Ee_i ($i \in \{1,2,...,5\}$), the HMM must continue by progressing to either Eb_{i+1} or Ee_{i+1} or by leaving the sheet states (and entering a helix, turn, or coil state). In other words, the HMM must progress through the sheet states in an ordered fashion (but may switch among the buried and exposed states) until such time as it leaves the sheet. Once in the states Eb_6 or Ee_6, the HMM can remain there (with the possibility of switching between these two states) or can leave the sheet. This arrangement of 12 sheet states is illustrated in Figure 2A. The manner of controlling the distributions of secondary structure lengths is described below.

The choice of six position-specific sheet states each for buried and exposed was made with a likelihood ratio testing procedure that took into account the number of parameters being estimated and the improvement in goodness-of-fit as the number of position-specific states increased. We use n_E to represent the number of position-specific states for each of buried and exposed sheet. Similar considerations suggested that the helix states Hb and He each be expanded into n_H = 10 states (Hb_1, Hb_2,..., Hb_{10} and He_1, He_2,..., He_{10}) and that the turn states Tb and Te be expanded to Tb_1, Tb_2, Te_1, and Te_2 ($n_T = 2$). Only two coil states (denoted Cb_1 and Ce_1) are in the HMM ($n_C = 1$) because coil represents a collection of diverse minor secondary structure elements (including but not limited to loops) that together have approximately a geometric length distribution. In summary, this led to 38 hidden states comprising 10 buried and 10 exposed helix states, six buried and six exposed sheet states, two buried and two exposed turn states, and one buried and one exposed coil state.

The model requires that all sites with a particular secondary structure and accessibility status experience the same amino acid replacement process, regardless of relative position within their secondary structure element. For example, a buried site that begins a helix (Hb_1) and a buried site in the third position of a helix (Hb_3) share the same replacement process. Thus, each of the 38 HMM states corresponds to a particular one of the eight amino acid replacement categories.

The most natural way to estimate transition probabilities (ρ_{ij}) among the 38 HMM states is to examine amino acid sequences of known structure, count how many times a site in state *i* is followed by a site in state *j*, and divide this count by the number of times sites in state *i* are followed by sites in any category. We felt there were insufficient data in the BRKALN database to make reliable estimates of all transition parameters in this manner. To reduce the number of parameters without a great sacrifice in the utility of the model, we make assumptions that we believe are reasonable albeit not strictly correct. Because these assumptions define the form of transition probabilities in our model, they are described here in detail and are illustrated in Figure 2.

If the current state is Xy_i (where X is one of H, E, T, C; *y* is *b* or *e*; and $i = 1,2,...,n_X$), the probability that the next state does not have secondary structure X is assumed to be independent of the current accessibility status (*y*). This probability is denoted g_i^X for $i = 1,2,...,n_X - 1$ (and X ≠ C) and $1 - r^X$ for $i = n_X$.

The probabilities g_i^X can be fixed so that the expected proportions of secondary structure X elements with given lengths between 1 and $n_X - 1$ can take any specified values. If the desired proportion of elements of length $i \in \{1,2,...,n_X - 1\}$ is f_i^X (and defining $f_0^X = 0$ for all X), then setting $g_i^X = f_i^X/(1 - \Sigma_{j=0}^{i-1} f_j^X)$ satisfies these length distribution requirements. We have estimated the f_i^X as the observed proportions of secondary structure elements of type X that have length exactly *i* in the BRKALN database and applied the above formula for the g_i^X.

The probabilities r^X determine the expected length distributions of secondary structures X for lengths $i \geq n_X$. Effectively, we model the tail of the length distribution with a geometric distribution. After selecting the value of n_X, we estimate the parameters r^X by the values that equate the observed mean lengths of secondary structure X elements in the BRKALN database and the expected mean lengths. We note that our estimates of g_i^X and r^X are maximum likelihood estimates for the case where the probabilites of lengths less than n_X are each represented by an individual parameter, and the probabilites of lengths n_X or greater are determined by attaching a geometric tail to the length distribution. A comparison between the observed and predicted distributions of secondary structure lengths is given in Figure 1, D–G.

TABLE 1

Estimated amino acid frequencies for the eight replacement categories

	Replacement category							
	H*b*	H*e*	E*b*	E*e*	T*b*	T*e*	C*b*	C*e*
A	15.3	10.9	7.4	3.0	8.1	5.7	6.5	5.7
R	3.6	7.5	3.0	7.0	4.6	4.2	2.8	5.3
N	2.9	4.4	1.6	2.1	4.2	6.9	3.2	5.0
D	2.2	7.2	2.4	5.0	6.7	8.7	6.0	8.7
C	2.8	0.5	4.6	0.7	3.3	0.5	2.6	0.6
Q	2.3	5.8	3.1	5.1	1.4	4.1	1.6	5.4
E	3.1	14.8	1.9	7.5	2.8	7.5	2.1	7.3
G	4.9	3.6	5.7	3.8	19.5	17.4	10.4	7.4
H	1.9	2.7	2.7	3.1	2.6	2.8	3.3	2.9
I	7.0	2.6	10.8	2.3	3.5	1.3	6.0	1.5
L	15.9	5.4	11.7	5.7	7.6	3.7	9.6	4.6
K	2.1	14.0	1.6	10.7	1.4	9.1	2.1	8.8
M	3.0	1.1	2.6	1.2	1.2	0.7	1.9	1.0
F	5.6	1.1	7.1	3.1	7.4	1.2	6.4	2.0
P	1.5	2.5	1.8	4.2	4.8	7.8	6.7	9.2
S	4.0	4.8	3.8	11.9	5.9	9.7	6.1	9.4
T	3.9	5.4	5.3	14.1	4.9	4.6	5.9	7.6
W	2.3	0.5	2.7	0.9	2.1	0.4	2.1	1.1
Y	4.5	2.0	7.7	2.6	2.9	1.3	5.7	2.5
V	11.1	3.2	12.5	5.9	5.1	2.4	9.1	3.8

Values listed are the parameters π_i^k as described in the text, given as percentages.

Given that the secondary structure W of the next state differs from the secondary structure X of the current state Xy_i, we assume that W and the next accessibility status $z = b$ or e are independent of the value of *i*. The conditional probabilities of transitions from Xy_i to Wz_1 are denoted $a_{Xy,Wz}$ and were estimated directly from the known structures in the BRKALN database by their observed relative frequencies.

Given that the secondary structure of the next state is identical to that of the current state Xy_i, we assume the accessibility status of the next state is independent of *i*. The probability that the next site is buried, given that the current site is buried and both sites have secondary structure X, is represented by p_b^X. The complementary probability $q_b^X = 1 - p_b^X$ is then the probability that the next site is exposed, given that the current site is buried and both sites have secondary structure X. Similarly, p_e^X is the probability that the next site is exposed, given that the current site is exposed and both are secondary structure X. We use q_e^X to represent the complementary probability $1 - p_e^X$. These probabilities are again estimated directly from the observed frequencies in the experimentally determined structures of the BRKALN database.

Relative rates of amino acid replacement: One way to compare the eight estimated amino acid replacement processes is to determine the relative rate at which sites in the different replacement categories evolve. We normalize rates so that the average site evolves at rate 1. First, note that the rate at which amino acid *i* is replaced in category *k* is $\Sigma_{j\neq i}\,\alpha_{ij}^k$, the sum of the replacement rates α_{ij}^k over all amino acids that are not *i*. Second, the overall replacement rate for category *k* is $\Sigma_i\pi_i^k\Sigma_{j\neq i}\,\alpha_{ij}^k$, which accounts for amino acid frequencies by weighting the amino acid-specific rates by π_i^k, the frequency of amino acid *i* in category *k*. Table 1 displays the estimated values of the π_i^k. Finally, different replacement categories *k* have different stationary probabilities Ψ_k, determined for our HMM by the equilibrium distribution of the matrix ρ_{ij}. Accounting for the variation in stationary probabilities among replacement categories, the relative rate of replacement for category *k* is

$$R^k = \frac{\Sigma_i\pi_i^k\Sigma_{j\neq i}\alpha_{ij}^k}{\Sigma_l\Psi_l\Sigma_i\pi_i^l\Sigma_{j\neq i}\alpha_{ij}^l}. \qquad (1)$$

Relative replacement rates associated with a particular secondary structure averaged over accessibility classifications can also be estimated. For example, the relative rate for helices is

$$\begin{aligned} R^{H\cdot} &= \frac{\Psi_{Hb}R^{Hb} + \Psi_{He}R^{He}}{\Psi_{Hb} + \Psi_{He}} \\ &= \frac{\Psi_{Hb}(\Sigma_i\pi_i^{Hb}\Sigma_{j\neq i}\alpha_{ij}^{Hb}) + \Psi_{He}(\Sigma_i\pi_i^{He}\Sigma_{j\neq i}\alpha_{ij}^{He})}{(\Psi_{Hb} + \Psi_{He})\Sigma_l\Psi_l\Sigma_i\pi_i^l\Sigma_{j\neq i}\alpha_{ij}^l} \qquad (2) \end{aligned}$$

Similarly, relative rates for each accessibility classification averaged over secondary structures (*e.g.*, $R^{\cdot b}$) can be estimated. These estimates are shown in Table 2.

A concern is that one or a few proteins in the BRKALN database could potentially have an especially large impact on the relative rate estimates. The potential for a small number of protein families to have a great impact exists because the rate estimates are based upon amino acid replacement counts from pairwise sequence comparisons. Protein families with many sites or many sequences will tend to generate higher counts that families consisting of a few short sequences. If these influential protein families tend to evolve via comparatively atypical evolutionary processes, the general applicability of our evolutionary model would be restricted. To investigate this, we employed a nonparametric bootstrap resampling procedure (Efron and Tibshirani 1993). One hundred resampled data sets were constructed by sampling 207 randomly selected protein families with replacement from the 207 families of the BRKALN database. From each resampled data set, stationary probabilities of the eight replacement categories for our model were estimated. Likewise, replacement counts were generated from each resampled data set. Each resampled data set thereby

TABLE 2

Relative rates of amino acid replacement

Structure	Buried	Exposed	Average
Helix	0.71 ± 0.08	1.57 ± 0.07	1.11 ± 0.06
	0.70 ± 0.09	1.54 ± 0.07	1.09 ± 0.07
Sheet	0.72 ± 0.05	1.20 ± 0.18	0.87 ± 0.04
	0.71 ± 0.06	1.18 ± 0.17	0.85 ± 0.04
Turn	0.56 ± 0.05	1.30 ± 0.09	1.08 ± 0.07
	0.55 ± 0.05	1.28 ± 0.08	1.06 ± 0.06
Coil	0.51 ± 0.03	1.23 ± 0.06	0.90 ± 0.04
	0.50 ± 0.03	1.35 ± 0.12	0.96 ± 0.06
Average	0.65 ± 0.04	1.35 ± 0.04	1
	0.64 ± 0.05	1.36 ± 0.05	1

For each entry in the table, two pairs of rate estimates and sample standard deviations are listed. The upper line of each entry refers to the case in which insertions relative to known structures are ignored. The lower line refers to the case in which insertions relative to known structures are classified as exposed coil.

yielded estimates of relative rates of replacement for each category and the sample standard deviations of these estimates were determined (see Table 2).

A decision that we had to make when analyzing the pairwise replacement count data was how to treat data from sites that had unknown secondary structure or accessibility because they were insertions relative to the experimentally determined structure. One possibility was to ignore these data. Another was to add these replacement count data to the exposed coil (C*e*) replacement category because of the observed tendency for insertion and deletion events to affect exposed coils. As the relative rate estimates in Table 2 exhibit, the treatments yield similar outcomes. All other results presented here are based on estimating rates by ignoring sites that are insertions relative to known structures.

Phylogenetic tree: Typically the form (topology and branch lengths) of the phylogenetic tree relating a set of sequences is unknown but of interest. In this case, it may be treated as a parameter of the problem and estimated using statistical methods. This is the approach that we adopt.

In protein structure prediction problems, it is now realized that evolutionarily related amino acid sequences should not be treated as though they are statistically independent of one another (Benner *et al.* 1994; Goldman *et al.* 1996). In fact, common ancestry induces a correlation structure among sequences that is specified by the phylogenetic tree relating them. It has been shown that ideas developed in comparative evolutionary biology (Felsenstein 1985; Harvey and Pagel 1991) for using phylogenetic trees to describe this correlation structure can improve the results of secondary structure prediction algorithms (Goldman *et al.* 1996). The model presented here provides a statistical basis for a protein secondary structure prediction technique that shares these benefits.

Likelihood calculations: After estimating the relative rates of replacement α_{ij}^k and the HMM transition probabilities ρ_{ij} from the BRKALN database, we can calculate the likelihood for a candidate phylogenetic tree T (representing both topology and branch lengths). We do not re-estimate the α_{ij}^k or ρ_{ij} during the likelihood calculations but instead fix them at their estimated values. We denote the aligned data set by S, its length (number of amino acids) by N, the first i columns of the data set by S_i, and the ith column itself by s_i. In the equations below, many of the probabilities are actually conditional upon α_{ij}^k and ρ_{ij} but, for the sake of clarity, we omit α_{ij}^k and ρ_{ij} when this is feasible. The likelihood of the tree T is given by $\Pr(S|T)$, and this is calculated via the terms $\Pr(S_i,c_i|T)$ for each possible secondary structure category c_i at site i using the iteration:

$$\Pr(S_i,c_i|T) = \sum_{c_{i-1}} \Pr(S_{i-1},c_{i-1}|T)\rho_{c_{i-1}c_i}\Pr(s_i|c_i,T) \qquad (3)$$

for $i > 1$. The terms $\Pr(s_i|c_i, T)$ are evaluated using the Markov process replacement models appropriate for each secondary structure c_i and the pruning algorithm of Felsenstein (1981). Because the site at the N terminus of a protein tends to be an exposed coil (Ce), we assume this is the case and start the iteration according to:

$$\Pr(S_1,c_1|T) = \Pr(s_1|c_1,T)\Pr(c_1|T) = \Pr(s_1|c_1,T)\delta_{c_1,Ce} \quad (4)$$

where $\delta_{c_1,Ce} = 1$ if $c_1 = Ce$, and 0 otherwise. This has proved slightly superior to using the stationary distribution (Ψ_i) to describe the secondary structure probabilities at the first site of sequences (results not shown), in which case $\Pr(S_1, c_1|T)$ is set equal to $\Pr(s_1|c_1,T)\Psi_{c_1}$. When completed, the iteration gives the required $\Pr(S|T)$ because

$$\Pr(S|T) = \sum_{c_N} \Pr(S_N,c_N|T). \qquad (5)$$

To obtain the maximum likelihood estimate $\hat{T}$, we use numerical optimization algorithms (*e.g.*, Swofford *et al.* 1996) to determine the T that maximizes $\Pr(S|T)$. A slight improvement in model fit compared to our treatment might be generated via special consideration of secondary structure elements at the extreme carboxyl-terminus of proteins. For example, column N of the alignment should not be the first position of an α-helix.

Prediction of protein secondary structure can subsequently be performed with a plug-in approach that fixes the tree at $\hat{T}$ and uses appropriate methods to calculate the posterior distribution $\Pr(c_i|S,\hat{T})$. This has been described in more detail for a simpler HMM of sequence evolution by Goldman *et al.* (1996) and will be explored in the future for the HMM introduced here.

Model comparisons: When a statistical comparison indicates that one evolutionary model is superior to a simpler model, this implies that features possessed by the complicated model and absent from the simple model may be evolutionarily important. This approach has been taken in the past to compare models of DNA sequence evolution (Goldman 1993; Yang *et al.* 1994, 1995), indicating the importance of factors such as base composition bias, transition/transversion rate ratio, and rate heterogeneity across DNA sites.

In this study, we investigate models that range from the simplicity of sites evolving identically to the relative complexity of the full version of our new HMM in order to test the significance of the different components of the new model. We introduce the following notation to label the model variants. If the number of different secondary structure types recognized is *ss*, the number of solvent accessibility classes distinguished is *acc*, and the total number of states in the HMM is *hmm*, then we can label models as $ss/acc/hmm^+$ ($^+$ indicating the use of the HMM to model dependencies between adjacent sites) or $ss/acc/hmm^-$ ($^-$ indicating the disabling of the HMM; see below). Using this notation, our new HMM is denoted $4/2/38^+$. The following models were also considered:

1/1/1⁻: This model represents the simplest case, in which no information on protein structure or nonindependence of sites is incorporated (hence, $ss = 1$ average secondary structure category, $acc = 1$ average accessibility state, and no meaningful HMM is possible). This corresponds to currently available methods implemented in software such as PAML (Yang 1997) and MOLPHY (Adachi and Hasegawa 1995). We have performed some analyses with each of three variants of this model: one derived from the work of Dayhoff *et al.* (1978) and denoted $1/1/1D^-$, one derived from Jones *et al.* (1992) and denoted $1/1/1J^-$, and the third derived from the BRKALN database of this study and denoted simply $1/1/1^-$.

$4/2/8^+$: This model adds to the $1/1/1^-$ model an HMM recognizing four secondary structure elements, each with two accessibility categories. No special allowance is made for the distributions of lengths of secondary structures (which therefore follow a purely geometric distribution under this model).

$4/1/19^+$: This model adds to the $1/1/1^-$ model an HMM recognizing four secondary structure elements and making allowance for the distributions of lengths of these structures but does not recognize different solvent accessibilities.

For all the above models, the methods for estimating model parameters varied slightly from those described above for the $4/2/38^+$ model because not all of the parameters of that model are meaningful for simpler versions. In all cases, methods analogous to those described above for the $4/2/38^+$ model were used to estimate appropriated parameters from the BRKALN database (but see above regarding the $1/1/1/D^-$ and $1/1/1J^-$ models); full details are available from the authors.

$4/2/38^-$: This model is very similar to the $4/2/38^+$ model,

but sites are treated as independent of one another. This independence is achieved by replacing each row of the matrix (ρ_{ij}) with that matrix's equilibrium distribution (Ψ_i).

As with the $4/2/38^+$ model (above), the $4/2/8^+$ and $4/2/38^-$ models are implemented with the first site in a protein forced to be an exposed coil. For the $4/1/19^+$ model, accessibility status is not considered, and we simply treat the first site as a coil.

ANALYSIS OF EXAMPLE DATA SETS

Data sets: We analyzed 11 data sets that encompass a broad range of genes, organisms, and evolutionary divergence. When possible, we utilized previously published amino acid sequence alignments to reduce the chances that results could be biased by our own alignment procedures or prejudices. In addition, we selected sequences that were sufficiently similar to be likely to share the same secondary structure.

Gapped positions in the alignments were treated as missing information, as they are in the maximum likelihood programs of the PHYLIP package (Felsenstein 1995). In regions where the alignment was deemed relatively unreliable, we treated the residues responsible for the alignment difficulty as missing information. This was done to reduce the impact of alignment errors on the evaluation of our models.

ADP-glucose pyrophosphorylase (abbreviated to ADPGP): Four plant sequences were used, as analyzed by Goldman and Yang (1994) in a study of improved models of DNA nucleotide evolution.

HIV-1 gp120 envelope glycoprotein (GP120): One sequence was selected from each of the eight major subtypes (A–H) of HIV-1.

HIV-1 p17 matrix protein: Sequences were selected from a number of strains of human immunodeficiency virus type 1 (HIV-1). Two data sets were studied, the first (P17ALL) comprising one sequence from each of the seven subtypes (A–D, F–H) of HIV-1 for which significantly different p17 sequences have been recognized and the second (P17B) comprising eight sequences from HIV-1 subtype B.

Sucrose synthase (SUSY): Four dicotyledonous plant sequences were used, as analyzed in a preliminary study of the effects on protein structure on sequence evolution by Thorne *et al.* (1996).

Xylanase (XYLA): Seven prokaryotic sequences were used, as analyzed in a preliminary study of phylogeny- and likelihood-based methods of protein secondary structure prediction by Goldman *et al.* (1996).

The following five data sets were derived from multiple sequence alignments deposited at the EMBL-European Bioinformatics Institute (EBI) and available electronically from ftp://ftp.ebi.ac.uk/pub/databases/embl/align. In each case, minor alterations were made by hand.

Alcohol dehydrogenase: Two data sets were formed from the alignment available from the EBI ftp server, file ds14642.dat (see also Yokoyama and Harry 1993). The first (ADHAN) was formed from 16 mammalian, avian, and amphibian sequences, with the homologous sequence from the cod *Gadus callarias* added by the authors. The second (ADHPL) comprises 13 plant sequences.

Glutamate dehydrogenase (GDH): Eleven sequences, with both eubacterial and eukaryotic representatives, were selected from the alignment file ds20281.dat (see also Teller *et al.* 1995).

G protein α subunit (GPA): Yokoyama and Starmer (1992) aligned a large number of G protein α subunit sequences (file ds15369.dat). We selected 18 $G_{i\alpha}$ and $G_{o\alpha}$ sequences from mammals, *Drosophila melanogaster*, and *Caenorhabditis elegans*.

Phosphoenolpyruvate carboxykinase (PEPCK): We have used the alignment of 18 Lepidopteran sequences studies and submitted (file ds24063.dat) by Friedlander *et al.* (1996).

Model comparisons: Table 3 contains maximum log-likelihoods obtained when analyzing each of the above 11 data sets with models ranging from the most simple ($1/1/1^-$, $1/1/1D^-$, $1/1/1J^-$) to the most complex ($4/2/38^+$). To better understand which specific features of our models are most responsible for improved fits, we performed a series of parametric bootstrap analyses on each data set. Each analysis was a comparison of a relatively simple model (the null hypothesis H_0) with a model that considers more aspects of protein structure (the alternative hypothesis H_A).

We use a likelihood-ratio test with test statistic Δl, the maximum log-likelihood for the alternative hypothesis minus the maximum log-likelihood for the null hypothesis. For the data sets analyzed here, this statistic can be computed from the appropriate entries in Table 3. To approximate the distribution of Δl under the null hypothesis, 100 simulated data sets are produced. The null model of sequence evolution is used, along with the maximum likelihood topology and branch lengths estimated under the null hypothesis for the original data set, to generate each simulated data set. The simulated data sets have the same number of taxa and are the same length as the original data set. If a residue at a particular site in the data set has unknown type in the original data set because of alignment uncertainty or gaps, the residue at this position is also considered unknown in the simulated data set. For each simulated data set, Δl can be calculated via likelihood maximization under the null and alternative hypotheses. If the observed value of Δl for the original data set is sufficiently extreme relative to the distribution of simulated values, then the null hypothesis can be rejected. One measure of extremity is given by the proportion of simulated test statistic values that equal or exceed the actual

TABLE 3

Maximum log-likelihoods for the analysis of 11 data sets under seven model variants

	Evolutionary model						
Data set	$1/1/1D^-$	$1/1/1J^-$	$1/1/1^-$	$4/2/8^+$	$4/1/19^+$	$4/2/38^-$	$4/2/38^+$
ADPGP	−2396.97	−2375.91	−2368.13	−2366.42	−2370.37	−2362.15	−2364.01
GP120	−3867.66	−3817.31	−3832.66	−3793.07	−3818.37	−3800.16	−3793.47
P17ALL	−1009.89	−990.04	−991.75	−985.50	−989.68	−985.40	−984.90
P17B	−632.59	−627.12	−629.46	−624.37	−626.32	−627.02	−624.36
SUSY	−4401.75	−4348.94	−4343.29	−4334.01	−4340.48	−4339.56	−4334.15
XYLA	−3162.47	−3144.63	−3127.90	−3093.86	−3117.37	−3096.39	−3092.16
ADHAN	−4366.52	−4340.12	−4317.73	−4263.53	−4299.18	−4265.68	−4261.37
ADHPL	−3002.63	−2993.03	−2977.90	−2949.91	−2971.18	−2945.51	−2948.48
GDH	−6930.06	−6354.10	−6307.92	−6177.30	−6279.40	−6181.02	−6176.34
GPA	−3184.37	−3154.48	−3159.05	−3113.00	−3141.32	−3122.38	−3112.17
PEPCK	−2699.73	−2702.36	−2686.75	−2615.31	−2675.13	−2609.25	−2615.28

value. This proportion is an estimate of the probability under H_0 of realizing a value of Δl at least as extreme as that observed for the original data. Sufficiently low values imply rejection of H_0. Another measure of extremity is a *z*-score, calculated by subtracting the mean simulated test statistic value from the actual value and then dividing by the sample standard deviation of the simulated values.

In our parametric bootstrap comparisons, we do not account for uncertainty in parameters governing the relative rates of amino acid replacement or organization of secondary structure and solvent accessibility. Estimates for these parameters are fixed at the values obtained from the BRKALN database. Only the topology and branch lengths are estimated from the simulated data sets. Ideally, the parametric bootstrap procedure would account for the uncertainty in all parameters when comparing models, but this would be computationally expensive.

Parametric bootstrap comparisons between some models are more informative than those between others. By combining evolution with protein structure, four potentially important features that we have incorporated in the $4/2/38^+$ model are (1) association of secondary structure and amino acid replacement dynamics, (2) association of solvent accessibility and replacement dynamics, (3) regional organization of secondary structure and solvent accessibility along a sequence, and (4) a relatively realistic length distribution of secondary structure elements. Earlier work (Thorne *et al.* 1996) indicated that the effect of secondary structure on amino acid replacement is important. Therefore, we have concentrated on parametric bootstrap analyses that assess the remaining three features. The comparison between a null hypothesis of the $4/1/19^+$ model and the $4/2/38^+$ model investigates the effect of solvent accessibility on amino acid replacement. The comparison between a null hypothesis of the $4/2/38^-$ model and the $4/2/38^+$ model addresses regional organization of secondary structure and solvent accessibility along a sequence. The comparison between a null hypothesis of the $4/2/8^+$ model and the $4/2/38^+$ model focuses on modelling realistic length distributions for secondary structure. We found it also of interest to use the $4/1/19^+$, $4/2/38^-$, $4/2/8^+$, and $4/2/38^+$ models as alternative hypotheses when the $1/1/1^-$ model was the null hypothesis. These comparisons explore how much various combinations of the aforementioned four features improve upon a model that neglects structure. The results of these parametric bootstrap comparisons are shown in Table 4.

Computation times: Because of the 38 HMM states and the eight replacement categories, our $4/2/38^+$ model is more computationally demanding than the $1/1/1^-$ model. Our experience is that the computation time required for analyzing data sets is more sensitive to the number of replacement categories than to the number of HMM states. For this reason, one might expect the $4/2/38^+$ model to require approximately eight times more computation than the $1/1/1^-$ model. In our experience, the actual ratio of CPU times required by these two models varies somewhat between analyses. As an example, in a case where topology was fixed and branch lengths were estimated, analysis of the GPA data set on a Digital Alphastation 500/400 required 199 seconds of CPU time with the $1/1/1^-$ model and 943 seconds with the $4/2/38^+$ model.

DISCUSSION

An earlier study (Thorne *et al.* 1996) that considered just three structure categories (α-helix, β-sheet, and loop) demonstrated that consideration of secondary structure can significantly improve the fit of models to data. The belief that secondary structure is important is reinforced by the results shown in Table 4. All but one of the comparisons in which the null hypothesis makes no distinction between secondary structures and the alternative employs different Markov models of amino acid replacement for different secondary struc-

TABLE 4

Results of parametric bootstrap analyses

Row	Hypotheses compared	ADPGP	GP120	P17ALL	P17B	SUSY	XYLA	ADHAN	ADHPL	GDH	GPA	PEPCK
(1)	H_0: 1/1/1$^-$	**4.12**	**39.19**	**6.85**	**5.10**	**9.14**	**35.74**	**56.36**	**29.42**	**131.58**	**46.88**	**71.46**
	H_A: 4/2/38$^+$	*3.1*	*12.2*	*4.8*	*4.2*	*4.7*	*9.9*	*11.7*	*9.4*	*22.4*	*13.0*	*16.9*
(2)	H_0: 1/1/1$^-$	−2.24	**14.29**	**2.07**	**3.14**	**2.82**	**10.53**	**18.54**	**6.72**	**28.52**	**17.73**	**11.62**
	H_A: 4/1/19$^+$	*0.4*	*6.8*	*2.9*	*3.1*	*2.6*	*4.9*	*6.0*	*4.3*	*8.6*	*7.6*	*5.1*
(3)	H_0: 1/1/1$^-$	**5.98**	**32.51**	**6.35**	**2.39**	**3.73**	**31.52**	**52.05**	**32.39**	**126.90**	**36.68**	**77.50**
	H_A: 4/2/38$^-$	*4.1*	*13.7*	*4.8*	*3.4*	*2.9*	*10.0*	*13.0*	*13.0*	*24.0*	*12.5*	*19.8*
(4)	H_0: 1/1/1$^-$	**1.71**	**39.59**	**6.25**	**5.09**	**9.28**	**34.04**	**54.20**	**28.00**	**130.62**	**46.05**	**71.44**
	H_A: 4/2/8$^+$	*2.3*	*12.2*	*4.5*	*4.3*	*4.9*	*9.7*	*11.5*	*9.3*	*21.8*	*13.2*	*17.0*
(5)	H_0: 4/1/19$^+$	**6.36**	**24.90**	**4.78**	**1.96**	**6.33**	**25.21**	**37.81**	**22.70**	**103.06**	**29.15**	**59.85**
	H_A: 4/2/38$^+$	*4.9*	*12.2*	*3.6*	*2.6*	*3.8*	*8.1*	*11.9*	*9.7*	*22.3*	*11.5*	*19.2*
(6)	H_0: 4/2/38$^-$	−1.86	**6.69**	0.50	**2.71**	**5.41**	**4.22**	**4.31**	−2.97	**4.68**	**10.20**	−6.04
	H_A: 4/2/38$^+$	*0.9*	*3.9*	*1.4*	*2.7*	*3.6*	*3.9*	*3.0*	*0.8*	*3.4*	*4.8*	*−0.2*
(7)	H_0: 4/2/8$^+$	**2.41**	−0.40	0.60	0.01	−0.14	**1.70**	**2.16**	**1.42**	**0.97**	0.83	0.02
	H_A: 4/2/38$^+$	*2.7*	*0.0*	*1.3*	*0.3*	*0.4*	*2.0*	*2.3*	*2.0*	*1.7*	*1.2*	*0.3*

As described in the text, seven different hypothesis tests (rows 1–7) were performed for each of 11 data sets. Two entries are listed for each combination of hypothesis test and data set. These entries are the following: (upper) the observed value of Δl, the log-likelihood for H_A minus the log-likelihood for H_0, and (lower) a *z*-score (italic type) that measures the estimated number of standard deviations by which the observed value of Δl exceeds the mean of the simulated values. When fewer than 5 of 100 simulated Δl values equal or exceed the observed value, the observed Δl value is shown in bold type.

tures strongly reject the null hypothesis (Table 4, rows 1–4). Further evidence for an association between secondary structure and amino acid replacement is summarized in Table 2. We note the surprising results that the estimated rates for buried helix and sheet residues are greater than those for buried turns and coils and that the estimated rate for exposed helix residues is greater than those for other exposed residues. Because the probability a site has a particular accessibility status is not independent of its secondary structure, these patterns are less clear for the weighted average over buried and exposed sites. Nevertheless, it is still contrary to intuition that α-helix and β-sheet positions experience more amino acid replacements on average than coil positions.

Differences in amino acid replacement dynamics associated with solvent accessibility status have been explored by Koshi and Goldstein (1995), but their significance to protein evolution has not been statistically tested. In our study, the tests represented in Table 4, rows 1, 3, 4, and 5, have a bearing on this question. Rows 1, 3, and 4 represent tests in which the effect of accessibility is studied in combination with other components of the models. The comparison between the 4/1/19$^+$ and 4/2/38$^+$ models (Table 4, row 5) directly investigates the effect of incorporating accessibility into our evolutionary model. For every data set and for all of these tests, the model incorporating solvent accessibility information is strongly preferred.

The comparison of the 4/2/38$^-$ and 4/2/38$^+$ models (Table 4, row 6) tests for evidence of correlation of structure categories along sequences. For some data sets, the null hypothesis of no correlation could not be rejected. This could reflect failure of the 4/2/38$^+$ model to incorporate information from residues that are not close in the linear sequence. Methods for detecting long-range interactions in an evolutionary context are currently being developed (D. D. Pollock, unpublished results), and we hope that the information they yield can be incorporated into models used for phylogenetic inference. An alternative explanation for failing to reject the null hypothesis of no correlation is that some data sets may have too few sequences or too little evolutionary divergence to provide relative certainty about underlying structural environments. It is difficult to detect regional organization of structure along a sequence if there is a high degree of uncertainty regarding the underlying structural environment at each site.

We have preliminary indications that secondary structure predictions generated from our 4/2/38$^+$ model are superior to those from the 4/2/38$^-$ model. For example, predictions of the 4/2/38$^-$ model exhibit an abundance of unrealistically short α-helices and β-sheets (results not shown). It may be the case that incorporating correlation of structure categories will be important for accurate secondary structure prediction, even for data sets where it has little impact on the goodness of fit of models. This possibility will be a focus of future research.

For six of the 11 data sets, we find that the comparison between the 4/2/8$^+$ and 4/2/38$^+$ models does not reject the former. For these data sets, there is no significant evidence in favor of the more complex HMM that attempts to use information regarding the distribution of lengths of secondary structure elements. Mixed results across different data sets (Table 4, row 7) again make it unclear whether this is a failing of our model or is due to insufficient data. Detection of both structural correlation along sequences and information pertaining to the length distributions of secondary structure elements might be assisted by adding more sequences to a data set. This should in effect increase the information pertaining to each alignment position and make the HMM states "less hidden."

Improved models of sequence evolution can lead to improved estimates of both evolutionary relationships (topology) and distances (branch lengths) in phylogenies (*e.g.*, Yang *et al.* 1994). Naylor and Brown (1997) have recently illustrated adverse effects on phylogenetic tree estimation from mammalian mitochondrial protein sequences when physicochemical properties of amino acids are ignored. We hope that our new ($4/2/38^+$) model may give more reliable phylogenetic estimates than simpler models. It directly incorporates four components representing structural features that have not generally been considered in models of sequence evolution, instead of using physicochemical properties of amino acids as surrogates. Two of these components, encompassing effects of secondary structure and accessibility, give particularly large improvements in the fit of models across the broad range of data sets that were studied.

One result of Naylor and Brown (1997) seems to contradict our finding that buried sites evolve more slowly than exposed sites. Naylor and Brown classified sites as hydrophilic or hydrophobic on the basis of the amino acids found in the alignment column at that site. In the data sets we analyzed, hydrophilic sites tend to be exposed to solvent and hydrophobic sites tend to be buried. With a parsimony analysis, Naylor and Brown (1997) concluded that hydrophilic sites fit an accepted tree topology better than hydrophobic sites. We attribute this difference in fit (as measured by parsimony techniques) to heterogeneity of rates among sites: slowly evolving sites generate less homoplasy than quickly evolving sites. Therefore, their results could be explained if, in contrast to our results, hydrophobic (buried) sites were evolving more quickly on average than hydrophilic (exposed) sites.

Fortunately, the two studies can be reconciled by realizing that our work is based on experimentally determined structures that are exclusively globular proteins whereas Naylor and Brown (1997) studied only integral membrane proteins. Relatively unconstrained sites on the surface of a globular protein are apt to be exposed to solvent and hydrophilic, but the relatively unconstrained sites on the surface of a membrane protein are likely to be lipid accessible and hydrophobic. The evolutionary processes of globular and membrane proteins are apt to differ, and Jones *et al.* (1994) have demonstrated that patterns of amino acid replacement in integral membrane proteins bear little resemblance to those in globular proteins.

In this article we have described the analysis of amino acid sequences under the assumption that the protein's true structure is unknown. In the case that the tertiary structure has been determined for a protein, the phylogenetic estimation method can be modified to allow the known secondary structure and accessibility information to be used. This would remove some of the sources of uncertainty and should improve the estimation of phylogenies. In this way, experimentally determined structures could directly assist phylogenetics instead of being ingored, or being used indirectly through their influence on average properties of databases of known structures as in this article.

Our amino acid replacement models for the eight structural environment categories are based only on analysis of database sequences. They are not varied to suit specific proteins. A promising approach for tailoring amino acid replacement processes to specific proteins has been developed by Cao *et al.* (1994) for an evolutionary model that ignores protein structure. This technique combines parameters estimated from database sequences with amino acid frequencies calculated from the specific sequences being analyzed and potentially could be extended to models that consider protein structure.

All models are potentially misled by violations of their assumptions. Our model assumes that all residues of an alignment are related via the same phylogeny, for example, that recombination is absent, but the HIV-1 data sets we have analyzed have potentially been subject to intra- and interspecific recombination (Robertson *et al.* 1995). Additionally, although structure is more conserved than sequence (*e.g.*, Chothia and Lesk 1986; Russell *et al.* 1997), it clearly does evolve. For example, there is some evidence that the secondary structure of one of the proteins we have examined (GP120) can vary in different strains of HIV-1 (Hansen *et al.* 1996). Our model currently assumes that there has been no change in protein secondary structure or accessibility since sequence divergence. Advanced models that explicitly address the evolution of structure would be of great interest for phylogenetic estimation, structure prediction, and the study of evolutionary processes. Although this would add complexity to our model, modern computational statistical methods may make such developments practical.

One of the most important advances in the reconstruction of evolutionary trees has been the consideration of heterogeneity of evolutionary rates among sequence sites (*e.g.*, Yang 1994, 1996). Although this variability has been statistically modelled, typically with a gamma-distribution, its biological basis has not been well characterized. Our estimates (Table 2) indicate that rate heterogeneity is strongly associated with structural environment. Exposed sites tend to experience ~2× the rate of amino acid replacements experienced by buried sites. A higher rate of replacement for exposed sites is seen for each secondary structure type. The association between accessibility status and replacement rates is a noteworthy feature of protein evolution but has received scant attention in the field of molecular evolution and has not been previously exploited in phylogenetic studies.

We believe that characterizing general associations of patterns and rates of sequence evolution with phenotypic features such as protein structure is essential to understanding the process of sequence evolution both within and between populations. It is the relationship between genotype and phenotype that drives much of evolution. Protein structure is fundamental to phenotype and yet little previous effort has been devoted to characterizing its impact on evolution.

Edward Holmes kindly assisted with the HIV-1 data sets. We also thank David Pollock, Pietro Liò, and two anonymous referees for their helpful suggestions. N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research. J.L.T. was supported by National Institutes of Health grant P01-GM45344 and the Alfred P. Sloan Foundation. D.T.J. is supported by a Royal Society University Research Fellowship. Software, parameter estimates, and data sets used for the analysis described above are available from the authors via http://ng-dec1.gen.cam.ac.uk/hmm/contents.html.

LITERATURE CITED

Adachi, J., and M. Hasegawa, 1995 *MOLPHY: Programs for Molecular Phylogenetics, Ver. 2.3.* Institute of Statistical Mathematics, Tokyo.

Asai, K., S. Hayamizu and K. Handa, 1993 Prediction of protein secondary structure by the hidden Markov model. CABIOS **9:** 141–146.

Benner, S. A., I. Badcoe, M. A. Cohen and D. L. Gerloff, 1994 *Bona fide* prediction of aspects of protein conformation: assigning interior and surface residues from patterns of variation and conservation in homologous sequences. J. Mol. Biol. **235:** 926–958.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice *et al.*, 1977 The protein data bank: a computer-based archival file for macromolecular structures. Eur. J. Biochem. **80:** 319–324.

Bleasby, A. J., and J. C. Wootton, 1990 Construction of validated, non-redundant composite protein sequence databases. Prot. Eng. **3:** 153–159.

Brown, M., R. Hughey, A. Krogh, I. S. Mian. K. Sjolander *et al.*, 1993 Using Dirichlet mixture priors to derive hidden Markov models for protein families, pp. 47–55 in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, edited by L. Hunter, D. Searls and J. Shavlik. AIII Press, Menlo Park, CA.

Bruno, W. J., 1996 Modelling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. **13:** 1368–1374.

Cao, Y., J. Adachi, A. Janke, S. Pääbo and M. Hasegawa, 1994 Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39:** 519–527.

Chothia, C., and A. M. Lesk, 1986 The relation between the divergence of sequence and structure in proteins. EMBO J. **5:** 823–826.

Churchill, G. A., 1989 Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51:** 79–94.

Dayhoff, M. O., R. V. Eck and C. M. Park, 1972 A model of evolutionary change in proteins, pp. 89–99 in *Atlas of Protein Sequence and Structure*, edited by M. O. Dayhoff. National Biomedical Research Foundation, Washington, DC.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt, 1978 A model of evolutionary change in proteins, pp. 345–352 in *Atlas of Protein Sequence and Structure*, edited by M. O. Dayhoff. National Biomedical Research Foundation, Washington, DC.

Efron, B., and R. J. Tibshirani, 1993 *An Introduction to the Bootstrap.* Chapman and Hall, New York.

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

Felsenstein, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

Felsenstein, J., 1995 *PHYLIP (Phylogenetic Inference Package), Ver. 3.57.* Department of Genetics, University of Washington, Seattle.

Felsenstein, J., and G. A. Churchill, 1996 A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13:** 93–104.

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., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Goldman, N., J. L. Thorne and D. T. Jones, 1996 Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263:** 196–208.

Gotoh, O., 1982 An improved algorithm for matching biological sequences. J. Mol. Biol. **162:** 705–708.

Hansen, J. E., O. Lund, J. O. Nielsen, S. Brunak and J.-E. S. Hansen, 1996 Prediction of the secondary structure of HIV-1 gp120. Proteins **25:** 1–11.

Harvey, P. H., and M. D. Pagel, 1991 *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford, UK.

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

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

Jones, D. T., W. R. Taylor and J. M. Thornton, 1994 A mutation data matrix for transmembrane proteins. FEBS Letts. **339:** 269–275.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kabsch, W., and C. Sander, 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers **22:** 2577–2637.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.

Koshi, J. M., and R. A. Goldstein, 1995 Context-dependent optimal substitution matrices. Prot. Eng. **8:** 641–645.

Lüthy, R., A. D. McLachlan and D. Eisenberg, 1991 Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins **10:** 229–239.

Naylor, G. J. P., and W. M. Brown, 1997 Structural biology and phylogeny estimation. Nature **388:** 527–528.

Overington, J., M. S. Johnson, A. Šali and T. L. Blundell, 1990 Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. Lond. Ser B **241:** 132–145.

Robertson, D. L., B. H. Hahn and P. M. Sharp, 1995 Recombination in AIDS viruses. J. Mol. Evol. **40:** 249–259.

Russell, R. B., M. A. S. Saqi, R. A. Sayle, P. A. Bates and M. J. E. Sternberg, 1997 Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J. Mol. Biol. **269:** 423–439.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis, 1996 Phylogenetic inference, pp. 407–514 in *Molecular Systematics*, Ed. 2, edited by D. M. Hillis, C. Moritz and B. K. Mable. Sinauer Associates, Sunderland, MA.

Taylor, W. R., 1988 A flexible method to align large numbers of biological sequences. J. Mol. Evol. **28:** 161–169.

Teller, J. K., P. J. Baker, K. L. Britton, P. C. Engel, D. W. Rice *et al.*, 1995 Correlation of intron-exon organisation with the three-dimensional structure in glutamate dehydrogenase. Biochim. Biophys. Acta **1247:** 231–238.

Thorne, J. L., N. Goldman and D. T. Jones, 1996 Combining protein evolution and secondary structure. Mol. Biol. Evol. **13:** 666–673.

Topham, C. M., A. McLeod, F. Eisenmenger, J. P. Overington, M. S. Johnson *et al.*, 1993 Fragment ranking in modelling of protein structure: conformationally constrained substitution tables. J. Mol. Biol. **229:** 194–220.

Wako, H., and T. L. Blundell, 1994 Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J. Mol. Biol. **238:** 682–692.

White, J. V., C. M. Stultz and T. F. Smith, 1994 Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. Math. Biosci. **119:** 35–75.

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

Yang, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993–1005.

Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analysis. TREE **11:** 367–372.

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

Yang, Z., N. Goldman and A. Friday, 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11:** 316–324.

Yang, Z., I. J. Lauder and H. J. Lin, 1995 Molecular evolution of the hepatitis B virus genome. J. Mol. Evol. **41:** 587–596.

Yokoyama, S., and W. T. Starmer, 1992 Phylogeny and evolutionary rates of G protein α subunit genes. J. Mol. Evol. **35:** 230–238.

Yokoyama, S., and D. E. Harry, 1993 Molecular phylogeny and evolutionary rates of alcohol dehydrogenases in vertebrates and plants. Mol. Biol. Evol. **10:** 1215–1226.

Communicating editor: G. B. Golding