Combining Protein Evolution and Secondary Structure

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

*Program in Statistical Genetics, Statistics Department, North Carolina State University; †Division of Mathematical Biology, National Institute for Medical Research, London; and ‡Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, London

An evolutionary model that combines protein secondary structure and amino acid replacement is introduced. It allows likelihood analysis of aligned protein sequences and does not require the underlying secondary (or tertiary) structures of these sequences to be known. One component of the model describes the organization of secondary structure along a protein sequence and another specifies the evolutionary process for each category of secondary structure. A database of proteins with known secondary structures is used to estimate model parameters representing these two components. Phylogeny, the third component of the model, can be estimated from the data set of interest. As an example, we employ our model to analyze a set of sucrose synthase sequences. For the evolution of sucrose synthase, a parametric bootstrap approach indicates that our model is statistically preferable to one that ignores secondary structure.

Introduction

It is widely recognized that evolutionary divergence of protein structures occurs much less rapidly than divergence of protein sequences (e.g., Chothia and Lesk 1986; Flores et al. 1993). This indicates that selective constraints may act to preserve protein structure. The nature of these constraints is poorly understood and they have received relatively little direct attention in the areas of population genetics and phylogeny reconstruction. We believe this lack of attention should be rectified. The process of molecular evolution will not be well understood until the constraints that affect it have been characterized.

There has been less work on modelling amino acid replacement than on modelling nucleotide substitution. This disparity may be attributable to the extra complexity of modelling the replacement process. Replacements tend to occur between chemically similar amino acid types and a replacement model should reflect this. There are many ways to categorize amino acids by chemical properties (e.g., hydrophobicity, charge, relative size of side chain), and physicochemical distances between amino acid types have been suggested (Grantham 1974; Taylor and Jones 1993), but these categorizations or physicochemical distances may not directly reflect the differences among amino acid types that are acted upon by evolution.

Cognizant of the difficulty of creating a realistic model for protein evolution based solely on physicochemical principles, Dayhoff and coworkers (Dayhoff, Eck, and Eck 1972; Dayhoff, Schwartz, and Orcutt 1978) developed an empirical approach. To construct their empirical amino acid transition matrix, sets of easily aligned (i.e., closely related) sequences were collected. Only closely related sequences were considered because, when evolutionary distance is sufficiently small, the possibility of multiple replacements can be ignored. The observed replacement patterns were used to construct a probabilistic replacement model. When the Dayhoff model was originally proposed, relatively few protein sequences were known. More recent studies (Gonnet, Cohen, and Benner 1992; Jones, Taylor, and Thornton 1992) followed the spirit of the Dayhoff approach but were able to tabulate a larger number of observed replacements. Because the Dayhoff model is empirical, it reflects the fact that different amino acid types are replaced at different rates and the fact that amino acids are usually replaced by chemically similar amino acids.

A problem with the Dayhoff approach is that it effectively models the replacement process at the “average” site in the “average” protein. There may be no such thing as an “average” site in an “average” protein. The physical environment of a protein site may greatly influence the replacement process at the site. Therefore, there may be variation among sites in the replacement process. Important features of the physical environment might include the secondary structure and whether the site is on the surface or in the interior of a protein. There have been tabulations of the observed number of amino acid replacements for each of several different categories of physical environment (e.g., Overington et al. 1990; Lüthy, McLachlan, and Eisenberg 1991; Topham et al. 1993; Wako and Blundell 1994), but the potential of these tabulations to disentangle phylogenetic correlations and constraints due to physical environment has not been exploited previously.

In this study, we introduce a probabilistic model that relates protein secondary structure to protein evolution. Our empirical approach is similar to the Dayhoff procedure but we model the replacement process for each of several categories of physical environment. We classify the physical environment of a site according to the secondary structure at the site. Three categories of secondary structure (α-helix, β-sheet, and loop) are be-
The Replacement Process for Each Category of Secondary Structure is Markovian and independent of the replacement process at other sites. This assumption is not ideal but is superior to completely ignoring secondary structure.

We utilize a data set maintained by one of us (D.T.J.) that contains representative sequences of 207 protein families. Protein families are only included in the data set if the tertiary structure of at least one member of the family has been experimentally determined. We will refer to this data set as the “known structure” data set. Most of the protein families in the “known structure” data set are represented by sequences in addition to the one with known tertiary structure. The members of each protein family are aligned. Fortunately, alignment is not difficult for these sequences because the sequences included in the data set are relatively similar to one another.

To convert information in the “known structure” data set from three-dimensional structure to secondary structure categories, we rely on the DSSP computer program (Kabsch and Sander 1983). This program can determine the secondary structure of a protein from the atomic coordinates that specify its tertiary structure. We proceed with the assumption that the category of underlying secondary structure is identical for residues of the column. Insertions and deletions tend to occur in regions of secondary structure that are loops (Benner and Gerloff 1991; Thornton et al. 1991). Therefore, we have chosen to classify these occasional columns in an alignment as belonging to the loop category.

The Replacement Process for Each Category of Secondary Structure

We model the replacement process for each category of secondary structure, whereas Dayhoff and coworkers (Dayhoff, Eck, and Eck 1972; Dayhoff, Schwartz, and Orcutt 1978; also see Kishino, Miyata, and Hasegawa 1990) model the process for an “average” site. We assume that, given the secondary structure at a site, the amino acid replacement process at the site is Markovian and independent of the replacement process at other sites. This assumption is not ideal but is superior to completely ignoring secondary structure.

Let \( p_k^j(t) \) be the probability, for secondary structure \( k \), of a site containing residue type \( j \) after an amount of evolution \( t \) given that the site initially contained residue type \( i \). If the amount of evolution separating two sequences is small, the possibility that multiple substitutions at a site have occurred since two sequences have diverged is negligible. In this case, \( p_k^j(t) \) can be approximated, when \( i \neq j \), by the product of \( t \) and the rate at which amino acid type \( i \) is replaced by amino acid type \( j \) for category \( k \) of secondary structure. This rate will be referred to as \( \alpha_{ij}^k \) and the array containing \( \alpha_{ij}^k \) for all combinations of amino acids and secondary structure categories will be referred to as \( \alpha \). Also, if \( t \) is small, \( p_k^j(t) \) is well approximated by

\[
1 - \sum_{j \neq i} \alpha_{ij}^k t.
\]

Consider a comparison between two protein sequences and assume \( N_m \) is the number of alignment positions where neither sequence has a gap. This study follows the example of Jones, Taylor, and Thornton (1992) in that pairwise comparisons are made only if the two sequences possess \( 0.85 N_m \) or more alignment positions with identical residues and at least one of the two proteins is the closest homolog in the data set to the other. The incentive for adopting the somewhat stringent sequence identity threshold is to make approximations for \( p_k^j(t) \) and \( p_k^j(t) \) accurate. The relationship between the threshold choice and resulting replacement model is discussed later in this paper.

Assume \( C \) is the number of sequence comparisons with similarity exceeding our threshold. Let \( n_{ij}^k \) be twice the number of alignment positions of category \( k \) in the \( C \) sequence comparisons that contained residues of type \( i \) in both sequences. Let \( n_{ij}^k \) be the number of alignment positions of category \( k \) where one of the two residues is type \( i \) and the other is type \( j \) (\( i \neq j \)). Notice that \( n_{ij}^k = n_{ji}^k \).

Our Markov model of replacement is constructed to be reversible. This means, if \( \psi_{ik}^j \) is the equilibrium probability of residue type \( i \) for category \( k \),

\[
\psi_{ik}^j p_k^j(t) = \psi_{jk}^i p_k^j(t) \forall \, i, j, t. \tag{1}
\]

Reversibility is ensured if

\[
\psi_{ik}^j \alpha_{ij}^k = \psi_{jk}^i \alpha_{ij}^k \forall \, i, j. \tag{2}
\]

Denoting the expected proportion of alignment positions in category \( k \) by \( \Psi_k \) and the amount of evolution separating the sequences in the \( m \)th comparison by \( t_m \), the reversibility property implies that for \( i \neq j \),

\[
E[n_{ij}^k] = \sum_{m=1}^{C} N_m \Psi_k (\psi_{ik}^j \alpha_{ij}^k + \psi_{jk}^i \alpha_{ij}^k) t_m
\]

\[
= 2 \Psi_k \psi_{ik}^j \alpha_{ij}^k \sum_{m=1}^{C} N_m t_m. \tag{3}
\]

Also,

\[
E\left[\sum_i n_{ij}^k\right] = 2 \Psi_k \psi_{ik}^j \sum_{m=1}^{C} N_m. \tag{4}
\]

If we approximate \( E[n_{ij}^k] \) by \( n_{ij}^k \) and \( E[\sum_i n_{ij}^k] \) by \( \Sigma p_k^j \), both of which can be observed from the “known structure” data set, then for \( i \neq j \),
The above equation shows we can estimate the relative rate of replacement from the product of $c_x$ and a constant $\alpha_x$. Applying this to equation (5), it can be shown that

$$\frac{n_{ij}}{\sum_j n_{ij}^3} = \frac{E[n_{ij}]}{\sum_j E[n_{ij}]} = \frac{\sum_m N_{m} n_{ij}^m}{\sum_m N_{m} \sum_i n_{ij}^m}.$$  

The above equation shows we can estimate the relative rate of replacement via the product of $c_x$ and a constant $\alpha_x$. Applying this to equation (5), it can be shown that

$$\frac{n_{ij}}{\sum_j n_{ij}^3} = \frac{E[n_{ij}]}{\sum_j E[n_{ij}]} = \frac{\sum_m N_{m} n_{ij}^m}{\sum_m N_{m} \sum_i n_{ij}^m}.$$  

The equation reveals that we can estimate the rate of change for category $k$ up to a constant. Applying this to the “known structure” data set, we find that, if the replacement rate of a loop site is scaled to be 1.0, then the replacement rate of an $\alpha$-helix site is 1.027 and the rate of a $\beta$-sheet site is 0.775.

It is interesting that the $\alpha$-helix rate estimate is greater than that for loops, but the biological significance is unclear. It may be the case that helices do evolve at greater rates than loops. Alternatively, this may be an artifact of our defining the loop category as consisting of all sites that are neither $\alpha$-helix nor $\beta$-sheet. Sites fitting this definition may actually be some mixture of those that are slowly evolving and those that are quickly evolving. We intend to investigate this issue in the future.

**Organization of Secondary Structure Along a Sequence**

The secondary structure at a site in a protein is not independent of secondary structure at surrounding sites; the types of secondary structure at adjacent sites are positively correlated. Secondary structure is a “regional” property of proteins. For this reason, we believe the organization of secondary structure along a sequence might be well described by a hidden Markov model.

Churchill (1989) was the first to apply hidden Markov models to a biological sequence. He employed hidden Markov models to describe the organization of AT-rich and GC-rich regions in DNA. Hidden Markov models have been used to describe the organization of secondary structure along a sequence (Asai, Hayamizu, and Handa 1993; White, Stultz, and Smith 1994), but these models were employed to analyze single protein sequences and phylogenetic correlations among sequences were therefore not an issue. The first use of hidden Markov models in sequence evolution is by Felsenstein and Churchill (1996) and has been implemented in versions 3.5 and later of the PHYLIP software package (see Fel-
Table 2
Estimated Probabilities of the Secondary Structure Category at a Position Given the Category at the Previous Position

<table>
<thead>
<tr>
<th>FROM CATEGORY</th>
<th>TO CATEGORY</th>
<th>α-helix</th>
<th>β-sheet</th>
<th>Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-helix ......</td>
<td>0.9085</td>
<td>0.0005</td>
<td>0.0910</td>
<td></td>
</tr>
<tr>
<td>(13,290)</td>
<td>(8)</td>
<td>(1.331)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-sheet ......</td>
<td>0.0051</td>
<td>0.8113</td>
<td>0.1836</td>
<td></td>
</tr>
<tr>
<td>(50)</td>
<td>(8,008)</td>
<td>(1.812)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loop ..........</td>
<td>0.0619</td>
<td>0.0862</td>
<td>0.8519</td>
<td></td>
</tr>
<tr>
<td>(1,341)</td>
<td>(1,867)</td>
<td>(18,450)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The observed number of transitions in the “known structure” data set between each pair of categories is in parentheses below the estimated probabilities of the transition.

senstein 1989). The Felsenstein–Churchill framework was developed to allow regional heterogeneity of nucleotide substitution rates but we modify it to permit regional heterogeneity of protein secondary structure.

With our hidden Markov model, we can analyze data sets consisting of aligned protein sequences with unknown structures. The hidden states of our model correspond to the underlying secondary structure. The secondary structure cannot be directly observed but it may be inferred through the pattern of amino acid replacements. Each category of secondary structure is associated with a distinct pattern of replacements and the task is to match the pattern of replacements with secondary structure.

Although the secondary structures of the sequences are unknown, we assume the organization of secondary structure along the sequences can be described by a stationary first-order Markov chain. This assumption implies that, given the secondary structure at site i, the secondary structure at site i+1 is independent of the secondary structures of sites 1, . . . , i−1. This assumption is biologically implausible. Secondary structure at a site is likely to be influenced not only by neighboring sites in the sequence but also by sites that are nearby at the level of tertiary structure. The model we adopt is not perfect but it is a starting point; other models of replacement completely ignore structure.

Sequences in the “known structure” data set can be used to estimate , the probability that site i has underlying secondary structure category c_i given that site i−1 has secondary structure category c_{i−1}. The matrix values of p_{c_{i−1},c_{i}} for each pair of secondary structure categories will be referred to as ρ. Imagine that c_i represents an α-helix and c_{i−1} represents a β-sheet. A simple way to estimate the probability that an α-helix position will follow a β-sheet position is to count the number of times for sequences of known tertiary structure that an α-helix position follows a β-sheet position and divide this number by the number of times that an α-helix or β-sheet or loop position follows a β-sheet position. Estimates obtained this way from the “known structure” data set are in table 2. Stationary probabilities of α-helix, β-sheet, and loop positions (i.e., the V_k values) derived from the transition probability estimates are respectively 0.325, 0.212, and 0.463.

Likelihood Calculations

We can use our model and a likelihood approach to reconstruct phylogenies from aligned protein sequences with unknown structures. The aligned data set will be assumed to have length N and will be represented by S. The first i columns of the aligned data set will be represented by S_i and the i_th column itself by s_i. To reconstruct a phylogeny, we want to estimate T, the topology and branch lengths of the tree. We estimate the rates of replacement (i.e., α) and the parameters for organization of secondary structure along the sequence (i.e., ρ) from the “known structure” data set. For the sake of simplicity, we will not introduce notation to distinguish between the true values of α and ρ and their estimates. Our goal is to find the T that maximizes Pr(S | T, α, ρ).

The first step is to calculate Pr(S_i | T, α, ρ) for some T of interest. This can be done with an iterative algorithm. At each step, the algorithm considers c_i, a possible secondary structure category for site i, and computes Pr(S_i | c_i | T, α, ρ). For i > 1, the algorithm uses the fact that

\[
Pr(S_i, c_i | T, α, ρ) = \sum_{c_{i-1}} Pr(S_{i-1}, c_{i-1} | T, α, ρ) \cdot Pr(c_i | S_{i-1}, c_{i-1}, T, α, ρ).
\]

The assumptions of our model allow the above to be simplified:

\[
Pr(S_i, c_i | T, α, ρ) = \sum_{c_{i-1}} Pr(S_{i-1}, c_{i-1} | T, α, ρ) \cdot Pr(c_i | S_{i-1}, c_{i-1}, T, α).
\]

To calculate Pr(s_i, c_i | T, α, ρ),

\[
Pr(s_i, c_i | T, α, ρ) = Pr(s_i | c_i, T, α, ρ) \cdot Pr(c_i | T, α, ρ)
\]

The pruning algorithm of Felsenstein (1981) can be applied to calculate Pr(s_i | c_i, T, α). The pruning algorithm requires determination of the replacement transition probabilities p_{c_i,c_j}(t) for each combination of amino acids and branch lengths. These transition probabilities can be calculated from α and t (e.g., see Goldman and Yang 1994). The iterative algorithm allows calculation of the likelihood because

\[
Pr(S | T, α, ρ) = \sum_{S_N} Pr(S_N, c_N | T, α, ρ).
\]

To estimate T, we use numerical optimization algorithms (e.g., Swofford et al., 1996) to find T, the value of T that maximizes Pr(S | T, α, ρ).

We can also use our model to reconstruct evolutionary trees in cases where a protein structure has been experimentally determined. In these cases, the amino
Table 3
Log-likelihoods of the Sucrose Synthase Data Set for Different Unrooted Tree Topologies and Replacement Models

<table>
<thead>
<tr>
<th>UNROOTED TOPOLOGY</th>
<th>MODEL</th>
<th>(Ath, Pau), (Stu, Vfa)</th>
<th>(Ath, Stu), (Pau, Vfa)</th>
<th>(Ath, Fva), (Pau, Stu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAY . . . .</td>
<td>−4,473.48</td>
<td>−4,401.75</td>
<td>−4,479.54</td>
<td></td>
</tr>
<tr>
<td>JTT . . . .</td>
<td>−4,422.01</td>
<td>−4,348.94</td>
<td>−4,426.12</td>
<td></td>
</tr>
<tr>
<td>THEM . . .</td>
<td>−4,417.19</td>
<td>−4,345.81</td>
<td>−4,420.76</td>
<td></td>
</tr>
<tr>
<td>US . . . .</td>
<td>−4,405.93</td>
<td>−4,339.35</td>
<td>−4,409.12</td>
<td></td>
</tr>
</tbody>
</table>

Note.—A list of the sources of the four sucrose synthase protein sequences along with abbreviations, common names, and Swiss-Prot accession numbers is: Arabidopsis thaliana (Ath, mouse-ear cress, Q00197), Phaseolus aureus (Pau, mung bean, Q01390), Solanum tuberosum (Stu, potato, P10691), and Vicia faba (Vfa, broad bean, P31926).

acid replacement process for each site is determined by the known secondary structure category at the site. Because the secondary structure is known, the likelihood calculations do not involve p. Instead, the likelihood is the product of likelihoods at individual sites, again calculated with the pruning algorithm of Felsenstein (1981). Reconstruction of phylogenies from sequences with known structures involves less uncertainty and is therefore expected to be more accurate than reconstruction of phylogenies from sequences with unknown structures.

Example and Parametric Bootstrap

To illustrate our approach, we consider four sucrose synthase protein sequences with unknown secondary structure. Sequences were aligned with the assistance of the Treealign software (Hein 1990). Regions of uncertain alignment may bias a phylogeny reconstruction procedure toward certain tree topologies. In columns where the alignment was deemed relatively uncertain, we treated all residues—or all but one residue—as residues of unknown amino acid type. This treatment of uncertain regions in the alignment should reduce the potential for alignment uncertainty to bias phylogeny reconstruction. Gap positions in the alignment were also treated as unknown residues. The resulting alignment contained 809 columns and a total of 19 residues of unknown amino acid type.

The four aligned sucrose synthase genes were analyzed by the replacement model described here and by three amino acid replacement models that ignore secondary structure. We will refer to the model described here as the US ("Uses Structure") model. The remaining three amino acid replacement models that we considered ignore secondary structure but otherwise employ approaches similar to that described above to estimate relative replacement rates from counts of observed replacements. The models referred to as DAY and JTT are respectively derived from the replacement counts of Dayhoff and collaborators (Dayhoff, Schwartz, and Orcutt 1978) and the replacement counts of Jones and collaborators (Jones, Taylor, and Thornton 1992). The model obtained from the "known structure" data set when protein secondary structure is ignored is referred to as THEM ("Typical Homogeneous Evolutionary Model").

Table 3 contains the maximum log-likelihood value for each combination of the three unrooted topologies and the four replacement models. The US model considers secondary structure and yields higher likelihoods than other models. The maximum-likelihood topology and branch length estimates obtained by this model are shown in figure 1. The cumulative CPU time required by a Sparcstation 10 Model 30 to estimate branch lengths of the three unrooted sucrose synthase topologies with the US model was 403.2 seconds. In general, the US model requires approximately the amount of computation required by THEM multiplied by the number of secondary structure categories.

Interestingly, the likelihoods of THEM for the sucrose synthase data are higher than those for the JTT or DAY models (table 3). One possible explanation is that THEM is based on replacement counts from the "known structure" data set. Protein families with known tertiary structure may not be a random sample of protein families (e.g., see Orengo, Jones, and Thornton 1994). For example, families of membrane-bound proteins are likely to be underrepresented in the "known structure" data set. This underrepresentation may be less severe in data sets used to construct the JTT or DAY models. Therefore, it may be the case that THEM is superior to the JTT and DAY models when some protein families are studied but is inferior when others are studied. Clearly, the potentially biased composition of the "known structure" data set could also affect the appropriateness of the US model in certain cases.

Although the US model produces higher likelihoods than THEM for the sucrose synthase data set, additional investigation is required before rejecting THEM in favor of the US model. The replacement process and secondary structure organization parameters of the US model were estimated independently of the su-
FIG. 2.—A histogram of simulated log-likelihood differences. The observed difference of 6.46 for the sucrose synthase data set is also indicated.

Sucrose synthase data; these parameters are not free to vary. Therefore, THEM is not a special case of the US model. Fortunately, the parametric bootstrap approach to model comparison described by Goldman (1993) does not require that models be nested. To employ the parametric bootstrap, we simulated 100 data sets according to our null hypothesis of THEM. We simulated under the topology and branch lengths that were estimated with THEM from the actual sucrose synthase data set.

Each simulated data set had the same alignment length and the same locations of residues of unknown type as the actual data set. For each simulated data set, we subtracted the maximum log likelihood of THEM from the maximum log-likelihood of the US model. If we calculate the mean and sample standard deviation of the 100 simulated differences, we find that the actual difference is 3.6 sample standard deviations higher than the mean of the simulated differences. A histogram of these 100 log-likelihood differences is shown in figure 2. The fact that each of the 100 simulated log-likelihood differences is less than the observed log-likelihood difference 6.46 indicates that we can reject THEM in favor of the US model. In other words, consideration of secondary structure appears to be warranted for the sucrose synthase data.

Discussion
Phylogeny Reconstruction

Because the US model is more biologically realistic than previously proposed models of protein evolution, it is reasonable to expect that it will yield better estimates of the phylogeny than other models (e.g., see Yang, Goldman, and Friday 1994). The statistical comparison of THEM and the US model for the sucrose synthase data adds support to this expectation but should not be viewed as a confirmation. Further study is necessary to verify that the US model is preferable to other models when the objective is phylogeny inference.

Bias and Variance of Replacement Rate Estimators

In this study, our empirically derived replacement model considers only sequence pairs from the “known structure” data set with 85% or more identical residues. If categories of secondary structure evolve at sufficiently different rates, this could potentially yield biased estimates of $\alpha_S^k$. Imagine two pairs of sequences, one rich in loop sites and the other rich in sheet sites, each separated by the amount of evolution $t$. If loop sites evolve more quickly than sheet sites, then the pair of sequences rich in loop sites may be less than 85% identical and the pair rich in sheet sites may be more than 85% identical. The pair rich in sheet sites but not the pair rich in loop sites would be considered in this study. This scenario could have effects that include loop sites being underrepresented.

Reliance upon pairwise comparisons in this study may increase the variance of $\alpha_S^k$ estimators. The same replacement event could be counted in several different pairwise comparisons. Imagine a group of four protein sequences and the unrooted topology that relates them. There are six possible pairwise comparisons between the four sequences. If all six comparisons were made and each comparison exceeded the 85% identity threshold, a replacement event that took place on the interior branch of the topology relating the four sequences would be counted in four of the six pairwise comparisons. In contrast, a replacement event that took place on any other branch of the topology would be counted in three of the six pairwise comparisons. To lessen this unbalanced counting of replacement events, the only pairwise comparisons that we make are those in which at least one of the sequences is being compared with its closest homolog in the data set. The unbalanced counting of replacement events that remains would inflate the variance of our estimators of $\alpha_S^k$ but would not bias them.

In the future, we plan to investigate estimators that avoid the potential bias and inflated variance described here. We also hope to determine the variances of these estimators.

Probabilistic versus Biological Assumptions

It is well known among molecular evolutionists that likelihood approaches are based on explicit probabilistic models but the fact that explicit probabilistic assumptions are not always easily translated to explicit biological meanings seems less appreciated. A likeli-
Acknowledgments

We will discuss this application in a forthcoming study. We hope our approach and its subsequent modifications will shed light on protein evolution and structure.

Future Directions

Modifications to our model can be envisioned and some have already been discussed above. Many of these modifications may improve the statistical performance of the model and some may be biologically meaningful. For example, heterogeneity of replacement rates among sites within a secondary structure category could be permitted by adopting approaches already developed for the purposes of modelling nucleotide substitution (see Yang 1993, 1995; Yang, Goldman, and Friday 1994; Felsenstein and Churchill, 1996). Another modification would be to increase the number of categories in our hidden Markov model. For example, sites on the surface of a protein could be distinguished from interior sites and the additional category “turn” could be included. Another research direction would involve combining DNA substitution models with our model of amino acid replacement. Recent work (Goldman and Yang 1994, Muse and Gaut 1994) indicates that this may be feasible.

The current version of our model is not tailored to any specific family of proteins but it may be possible to design models appropriate for specific families or even for particular sites within a specific family. Some protein families are particularly rich in α-helix positions and others in β-sheet positions (Levitt and Chothia 1976). Our hidden Markov model for organization of secondary structure along a sequence allows for “chance” variation in frequencies of secondary structure categories among protein families but the actual variation among protein families is probably larger than our model predicts. Perhaps this additional variation can be considered in future investigations. One concern that arises when such models are contemplated is whether sufficient information exists to estimate model parameters accurately.

It has not escaped our notice that this approach also has value in prediction of protein secondary structure. We will discuss this application in a forthcoming study. We hope our approach and its subsequent modifications will shed light on protein evolution and structure.

Acknowledgments

We thank Hirohisa Kishino, Joseph Felsenstein, and two anonymous reviewers for their constructive comments. We also thank Francis Giesbrecht, Elizabeth Thompson, and Simon Tavaré for their assistance with the quotation by W. G. Hunter. The research was supported by NIH grant R01 GM 49654, principal investigator J. L. Thorne. Programs written in Fortran that implement the techniques described here are available from N. Goldman (N.Goldman@gen.cam.ac.uk) and programs written in C are available from J. Thorne (thorne@stat.ncsu.edu).

LITERATURE CITED


Paul M. Sharp, reviewing editor

Accepted January 30, 1996