J. Mol. Biol. (1996) **263**, 196–208

JMB

Using Evolutionary Trees in Protein Secondary Structure Prediction and Other Comparative Sequence Analyses

Nick Goldman[1]*, Jeffrey L. Thorne[2] and David T. Jones[3]

[1]*Department of Genetics*
University of Cambridge
Downing Street, Cambridge
CB2 3EH, UK

[2]*Program in Statistical*
Genetics, Department of
Statistics, Box 8203, North
Carolina State University
Raleigh, NC 27695-8203
USA

[3]*Department of Biological*
Sciences, University of
Warwick, Coventry
CV4 7AL, UK

**Corresponding author*

Previously proposed methods for protein secondary structure prediction from multiple sequence alignments do not efficiently extract the evolutionary information that these alignments contain. The predictions of these methods are less accurate than they could be, because of their failure to consider explicitly the phylogenetic tree that relates aligned protein sequences. As an alternative, we present a hidden Markov model approach to secondary structure prediction that more fully uses the evolutionary information contained in protein sequence alignments. A representative example is presented, and three experiments are performed that illustrate how the appropriate representation of evolutionary relatedness can improve inferences. We explain why similar improvement can be expected in other secondary structure prediction methods and indeed any comparative sequence analysis method.

Keywords: comparative sequence analysis; evolution; hidden Markov model; phylogenetic trees; protein secondary structure prediction

Introduction

Numerous methods have been proposed that use multiple sequence alignments to assist in secondary structure analysis and prediction (e.g. Crawford *et al*., 1987; Zvelebil *et al*., 1987; Levin & Garnier, 1988; Rost & Sander, 1993; Stultz *et al*., 1993; Benner *et al*., 1994; Wako & Blundell, 1994b; Salamov & Solovyev, 1995). These methods have, however, used naive averaging, or at best highly *ad hoc* weighting schemes, to combine the contributions of all sequences. With unweighted averaging, the sequences are in effect being treated as though they were independent realizations of some process. Relationships between *ad hoc* weighting schemes and implicit assumptions about the process that generated the protein sequence data are less clear. Whatever the implicit assumptions are, there is no reason to expect that they are by chance either biologically meaningful or plausible.

It is appropriate to determine sequences' contributions to analyses in a manner that correctly represents the best estimate of the independent information they contain. For example, consider two homologous protein sequences. If these are closely related in evolutionary terms, they have had little time to diverge and will be quite similar. As a consequence, there will be little additional information in the second sequence that is not contained in the first. But if these sequences are more distantly related, they will be less similar and we expect more information to be available with the inclusion of the second. This has been recognized previously and has motivated heuristic algorithms used in structure prediction (e.g. see Benner *et al*., 1994) and weighting schemes in automated predictors that attempt to account for evolutionary conservation (e.g. see Rost & Sander, 1994). Other sequence weighting schemes have been proposed for sequences related by evolutionary trees (Altschul *et al*., 1989; Gerstein *et al*., 1994), but have not been applied to structure prediction. None of these methods has addressed the problem of non-independence *via* careful consideration of the statistical issues raised by evolutionary relationships. As a consequence, they extract less information pertaining to secondary structure than could be obtained by a more rigorous treatment.

The realization that there is non-independence of biological characteristics between study units (typically species) has already revolutionized studies in comparative evolutionary biology (e.g.

Abbreviations used: ML, maximum likelihood; HMM, hidden Markov model; PDB, Brookhaven Protein Data Bank.

Felsenstein, 1985; Harvey & Pagel, 1991), where the phrase ''phylogenetic inertia'' (Harvey & Purvis, 1991) is used to describe the non-independence of characteristics through descent from common ancestors. Analogous methodological advances are beginning to make inroads in population genetics (Felsenstein, 1992). There is no reason why the same ideas cannot be applied to the comparative analysis of amino acid and DNA sequences, where a large amount of research has already been performed on the estimation of phylogeny. The usual view of phylogenies as representing evolutionary relationships is valid but it is also valid to interpret phylogenies as representing the correlation structure among sequences. We explain here how prediction of protein secondary structure can be improved by careful consideration of phylogenetic relationships. Similar improvements can be expected in other comparative sequence analyses that do not consider phylogeny explicitly. For example, improvement can be anticipated in database searching algorithms.

When analyzing homologous sequences or other correlated data, methods that pay attention to the processes that generated the data are among the most powerful. By assuming and using an explicit probabilistic model for the evolutionary processes that generated the sequence data, homologous sequences can be analyzed with a maximum likelihood approach.

The standard maximum likelihood (ML) analysis for estimation of phylogenetic trees and branch lengths from aligned DNA sequences was originally described by Felsenstein (1981). Kishino *et al*. (1990) subsequently adapted the methodology to analysis of aligned protein sequences. With these methods, Markov process models of the nucleotide substitution or amino acid replacement processes are used to define probabilities, as functions of time, of any possible substitution or replacement. These probabilities, along with a phylogenetic tree topology and its branch lengths, define the likelihood function of the observed data for a candidate tree. The topology and associated branch lengths that maximize the likelihood are the ML estimates. This widely used procedure has been fully described by, for example, Felsenstein (1981) and Swofford *et al*. (1996).

Virtually all ML approaches are based on models of sequence evolution that assume that each position of the sequence alignment is independent of all others. Clearly, this is biologically unrealistic. For protein-coding sequences, the evolution of each site will be influenced by both neighboring and distant sites.

To overcome some of the limitations of ''independent sites'' models, and to illustrate the complete use of evolutionary information rather than crude averaging, we adopt a hidden Markov model (HMM) approach for secondary structure prediction. HMM approaches were introduced to the analysis of molecular sequences by Churchill (1989), who analyzed regions of varying G + C content in single DNA sequences. We use Churchill's HMM terminology. HMM approaches have recently been employed for the detection and characterization of sequence similarities (Baldi *et al*., 1994; Krogh *et al*., 1994) and have been successful despite the fact that these studies ignore effects of common ancestry. HMM applications in secondary structure prediction have only been to methods analyzing single sequences (Asai *et al*., 1993; White *et al*., 1994). HMM approaches in molecular evolution have been to model regional heterogeneity of substitution rate in aligned DNA sequences (Yang, 1995; Felsenstein & Churchill, 1996).

As with previous HMM approaches to secondary structure prediction, we categorize the secondary structure at each position of a sequence and assume that the organization of these categories along the sequence can be described by a Markov chain. The model is ''hidden'' because the secondary structure is not observed in the data (although it can be estimated). A Markov process model for each category of secondary structure describes evolution by amino acid replacement on the branches of a phylogenetic tree. These models differ between categories, reflecting differences in amino acid composition and in patterns and rates of amino acid replacement observed in different secondary structures. This contrasts with previous Markov process models of DNA substitution (e.g. see Jukes & Cantor, 1969; Felsenstein, 1981) and amino acid replacement (e.g. see Dayhoff *et al*., 1972, 1978; Kishino *et al*., 1990) because these attempt to describe the evolutionary process for an ''average'' site. Descriptions of amino acid replacement that differ according to context have been given in the past (e.g. see Topham *et al*., 1993; Wako & Blundell, 1994a,b; Koshi & Goldstein, 1995), but the potential for these descriptions to work in concert with an evolutionary tree has not been exploited. Our model is amenable to ML analysis, permitting estimation of phylogenetic relationships (evolutionary tree topology), evolutionary distances (branch lengths), and the secondary structure of the protein sequences analyzed.

We present an example analysis of the *Pseudomonas fluorescens* (PSEFL) xylanase sequence (known to adopt the TIM-barrel topology) and six homologs, and devise three experiments that illustrate how secondary structure prediction can be aided by the use of sequences homologous to the target sequence. More importantly, the experiments demonstrate the significant further improvement possible when the phylogenetic relationships of the sequences are considered. The application of our HMM to phylogenetic estimation has been described by Thorne *et al*., (1996).

Theory

Organization of secondary structure along sequences

The description of the organization of secondary structure along amino acid sequences defines the system equations of the HMM. We assume that c_i, the secondary structure at site i, depends only on c_{i-1}, the structure at site $i - 1$ (i.e. the site adjacent to i toward the amino-terminal end). This dependence is described by a set of stationary transition probabilities. We write the probability that a site is in category l if the preceding site is in category k as ρ_{kl}; clearly $\Sigma_l \rho_{kl} = 1$ for all k. Here, we have used three secondary structure categories: α-helix (α), β-sheet (β) and loop (L). The term loop is used solely to indicate that a site is in neither an α-helix nor a β-sheet.

A simple way to estimate the values of the ρ_{kl} is to examine amino acid sequences of known secondary structure, count the number of times sites in secondary structure category k are followed by sites in category l and divide this by the number of times sites in category k are followed by sites in any category. We have performed this estimation on sequences taken from the BRKALN database of structure-related multiple sequence alignments maintained by one of us (D.T.J., unpublished results). This database contains amino acid sequences classified into protein families for which the tertiary structure of at least one member has been experimentally determined. It was built by extracting a set of non-homologous protein sequences from the January 1995 release of the Brookhaven Protein Data Bank (PDB; Bernstein *et al*., 1977). In this instance 207 chains were extracted, no pair of which showed sequence identity greater than 25%. Low-resolution structures (resolution >2.6 Å) were excluded from consideration, along with NMR structures. For each chain from PDB, a dynamic programming-based sequence similarity search (Gotoh, 1982) was made through release 25.0 of the OWL non-redundant protein sequence database (Bleasby & Wootton, 1990) to find sequences with greater than 30% sequence identity with the sequence of the known structure. A multiple sequence alignment method (Taylor, 1988) was then applied to each of the resulting 207 families of sequences.

Secondary structure assignments were extrapolated across each aligned sequence family in BRKALN from the assignments for the master structure determined using the DSSP program (Kabsch & Sander, 1983). Where an insertion occurred in the middle of a secondary structure in the known structure, the inserted residues were given the same secondary structure assignment. This occurrence was rare, however, and mostly occurred as a result of probable β-bulges in some families. Insertions occurring elsewhere were defined to have unknown secondary structure and were excluded from all calculations of observed replacement counts. We refer to the resulting database as the DTJ-database. Both it and the BRKALN database are available by anonymous ftp from the directory pub/HMM at ftp.biochem.ucl.ac.uk.

Table 1. HMM of organization of secondary structure along sequences

	To		
From	α	β	L
α	0.9085	0.0005	0.0910
	(13290)	(8)	(1331)
β	0.0051	0.8813	0.1836
	(50)	(8008)	(1812)
L	0.0619	0.0862	0.8519
	(1341)	(1867)	(18450)
Ψ	0.3248	0.2124	0.4628

The Table shows the probability of a residue in a given secondary structure category being followed by a residue in each category, for the three categories considered in this work. Underneath, in parentheses, are the numbers of times each transition was observed in the sequences from the DTJ-database. The row labelled Ψ gives the resulting equilibrium distribution of secondary structure categories.

Only one sequence from each family was used in the estimation of the ρ_{kl}, to avoid biasing the estimates towards the patterns of secondary structure organization found in protein families that happen to have many members. We further classified the DSSP secondary structure assignments as follows: H, α-helix; A, E and P, β-sheet; all others, loop. Counts, estimates of ρ_{kl} and the resulting equilibrium distribution of secondary structure categories (Ψ_k) are shown in Table 1.

Amino acid replacement processes

The description of the evolutionary replacement of amino acids provides the observation equations of the HMM. We use different reversible Markov process models, each with 20 states corresponding to the 20 amino acids, for each of the three secondary structure categories in our model. To define the Markov process amino acid replacement model for each secondary structure category k, we must estimate $q_{ij}^{(k)}$, the instantaneous rate of replacement of amino acid i by amino acid j ($j \neq i$) for a site in secondary structure category k (we use parentheses about the k to distinguish it from an exponent). We write the matrix of values $q_{ij}^{(k)}$ as $\mathbf{Q}^{(k)}$, with equilibrium distribution $\boldsymbol{\pi}^{(k)}$ having elements $\pi_i^{(k)}$. The collection of rate matrices for all secondary structure categories will be represented by $\mathbf{Q}$. Reversibility then requires:

$$\pi_i^{(k)} q_{ij}^{(k)} = \pi_j^{(k)} q_{ji}^{(k)} \quad \forall i, j, k$$

where, by definition,

$$q_{ii}^{(k)} = -\Sigma_{j \neq i} q_{ij}^{(k)}.$$

We estimate the $\mathbf{Q}^{(k)}$ as follows (see also Jones *et al*., 1992). For each sequence of every family in the DTJ-database, the pair formed with its closest neighbor is considered. Pairs that after alignment have identical residues at less than 85% of alignment positions are discarded. This is to ensure that only evolutionarily closely related sequence pairs are used, a necessary condition for an approximation made below. Suppose now that the remaining sequence pairs are indexed

$m = 1, \dots, M$, and that pair m has N_m alignment positions where neither sequence has a gap. Let $n_{ij}^{(k)}$ be the total number of alignment positions of category k in the M comparisons where one of the two residues is amino acid i and the other is j, and let $n_{ii}^{(k)}$ be twice the total number of positions containing amino acid i in both sequences. Denoting the amount of evolution separating the sequences in comparison m by t_m, and with the equilibrium distribution of secondary structure categories Ψ_k as above, then:

$$\mathrm{E}[n_{ij}^{(k)}] = \sum_{m=1}^{M} \Psi_k N_m (\pi_i^{(k)} q_{ij}^{(k)} t_m + \pi_j^{(k)} q_{ji}^{(k)} t_m)$$

$$= 2\Psi_k \pi_i^{(k)} q_{ij}^{(k)} \sum_{m=1}^{M} N_m t_m \quad \forall\, i \neq j \tag{1}$$

This uses the fact that the probability of a replacement is approximately equal to the product of instantaneous rate of replacement ($q_{ij}^{(k)}$) and time (t_m) if the time is small and the probability of multiple replacements is negligible; hence the strict 85% sequence identity rule above. Also:

$$\mathrm{E}\left[\sum_r n_{ir}^{(k)}\right] = 2\Psi_k \pi_i^{(k)} \sum_{m=1}^{M} N_m \quad \forall i, r \tag{2}$$

If we approximate $\mathrm{E}[n_{ij}^{(k)}]$ by $n_{ij}^{(k)}$ and $\mathrm{E}[\Sigma_r n_{ir}^{(k)}]$ by $\Sigma_r n_{ir}^{(k)}$ then, for $i \neq j$:

$$\frac{n_{ij}^{(k)}}{\sum_r n_{ir}^{(k)}} \doteq \frac{\mathrm{E}[n_{ij}^{(k)}]}{\mathrm{E}\left[\sum_r n_{ir}^{(k)}\right]} = q_{ij}^{(k)} \frac{\sum_{m=1}^{M} N_m t_m}{\sum_{m=1}^{M} N_m} \tag{3}$$

This indicates that we can estimate all the rates of replacement ($q_{ij}^{(k)}$) to within an unimportant multiplicative constant that does not depend on i, j or k. The equilibrium frequencies of amino acids for each secondary structure category, $\boldsymbol{\pi}^{(k)}$, are determined by the $\mathbf{Q}^{(k)}$.

Since our rates of replacement are determined only to within a multiplicative constant, we can scale them so that the overall rate of replacement at equilibrium at a loop site, $\Sigma_i \Sigma_{j \neq i} \pi_i^{(L)} q_{ij}^{(L)} \equiv -\Sigma_i \pi_i^{(L)} q_{ii}^{(L)}$, equals 1. With this scaling, we find that the replacement rate at an α-helix site is 1.027, and that at a β-sheet site is 0.775. Figure 1 gives a graphical representation of our estimates of the scaled $\mathbf{Q}^{(k)}$ and of $\boldsymbol{\pi}^{(k)}$; precise values are available in electronic form upon request from the authors. It is of interest to note that the estimates from the DTJ-database of the replacement rates and equilibrium frequencies vary among secondary structure categories, supporting our expectation that evolution and secondary structure are not independent.

We scale time so that the lengths of branches of phylogenetic trees are measured in units of expected number of replacements per 100 sites, analogous to the PAMs of Dayhoff *et al.* (1972, 1978), except that our measure represents the expectation over three secondary structure categories.

An alternative method for estimating what are in effect our $\mathbf{Q}^{(k)}$ is given by Koshi & Goldstein (1995). Our results (Figure 1) may be compared with theirs (Koshi & Goldstein, 1995) and with those of Thompson & Goldstein (1996a).

Likelihood and probability calculations

We can use our model to estimate both phylogeny and secondary structure. If either is known with certainty, it may be incorporated into the model as a fixed value and should improve estimates of the other. Here, we describe the theory assuming both phylogeny and secondary structure are unknown and to be estimated. Suppose an aligned data set S is of length N, and let the first i sites be represented by S_i and the ith column itself be represented by s_i. The first stage is to estimate T, the topology and branch lengths of the evolutionary tree. Using an ML approach, we aim to find the T that maximizes $\Pr(S|T, \rho, \mathbf{Q})$. Since our estimates of ρ and $\mathbf{Q}$ are already determined (see above) from the DTJ-database, we drop them from the notation here.

It is possible to calculate $\Pr(S|T) \equiv \Pr(S|T, \rho, \mathbf{Q})$ recursively. For site i we compute $\Pr(S_i, c_i|T)$ for each possible secondary structure category c_i using the fact that, for $i > 1$:

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

$$\times \Pr(c_i|S_{i-1}, c_{i-1}, T)\Pr(s_i|S_{i-1}, c_{i-1}, c_i, T) \tag{4}$$

Due to the assumptions of our model, this can be simplified to:

$$\Pr(S_i, c_i|T) = \sum_{c_{i-1}} \Pr(S_{i-1}, c_{i-1}|T)\rho_{c_{i-1}c_i}\Pr(s_i|c_i, T) \tag{5}$$

Calculation of the terms $\Pr(s_i|c_i, T)$ is straightforward using the Markov process model of amino acid replacement developed above. The methods are essentially the same as those described by Felsenstein (1981) to calculate $\Pr(s_i|T)$ for a four-state model of DNA substitution. To calculate $\Pr(s_i|c_i, T)$ for our model, it is simply necessary to repeat this procedure independently for each secondary structure category.

Insertions and deletions are not accounted for in our model. To allow for analysis of alignments with indels, sequence positions with gaps are treated as unknown amino acids and make no contribution to overall likelihood calculations. In effect, the evolutionary tree for each alignment position is "pruned" so that only those branches leading to sequences with an amino acid at that position remain. This is the treatment of gaps used by the ML phylogeny inference programs in the PHYLIP package (Felsenstein, 1995). For our application, we

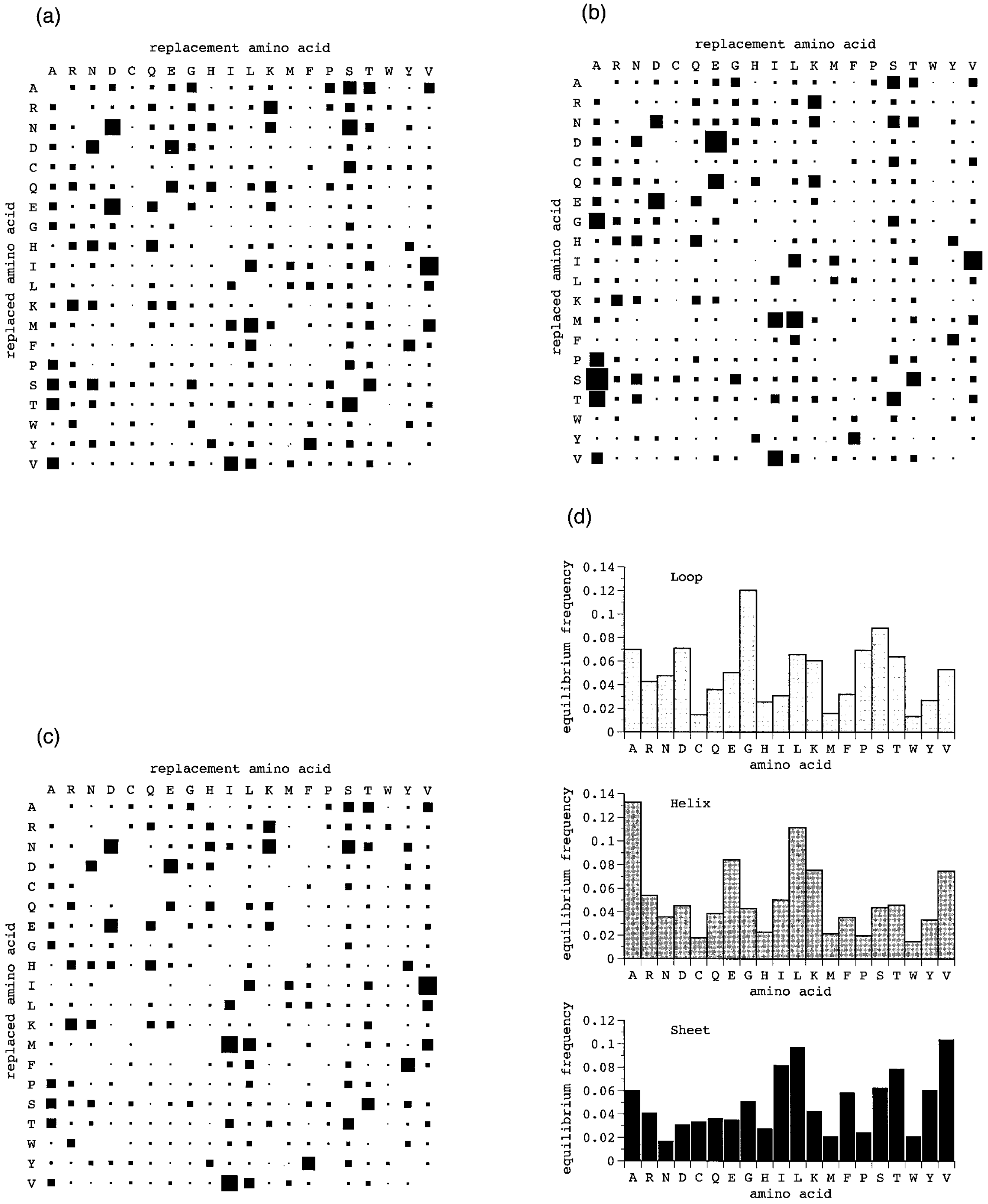

Figure 1. Graphical representation of amino acid replacement model parameters $\mathbf{Q}^{(k)}$ and $\pi^{(k)}$. (a) Areas of squares are proportional to the estimated replacement rates in matrix $\mathbf{Q}^{(\mathrm{L})}$. (b) $\mathbf{Q}^{(\alpha)}$. (c) $\mathbf{Q}^{(\beta)}$. Rates vary from 0 (e.g. $q^{(\mathrm{L})}_{\mathrm{NW}}$) to 0.834 ($q^{(\alpha)}_{\mathrm{SA}}$) when scaled as described in the text. (d) Heights of columns represent elements of the vectors of equilibrium distributions $\pi^{(\mathrm{L})}$, $\pi^{(\alpha)}$ and $\pi^{(\beta)}$.

think that this is preferable to the common practice in phylogenetics of deleting alignment positions where one or more sequences have gaps, because such a technique would disrupt information relating to organization of secondary structure along the sequences.

To start the recursion for $\Pr(S|T)$, note that:

$$\Pr(S_1, c_1|T) = \Pr(s_1|c_1, T)\Pr(c_1|T) = \Pr(s_1|c_1, T)\Psi_{c_1} \quad (6)$$

and to complete the calculation note that:

$$\Pr(S|T) = \sum_{c_N} \Pr(S, c_N|T) = \sum_{c_N} \Pr(S_N, c_N|T) \quad (7)$$

The ML estimate $\hat{T}$, the value of T that maximizes $\Pr(S|T)$, may be found with numerical optimization routines. These procedures are further described by Thorne *et al*. (1996).

Having found $\hat{T}$, it is further possible to estimate the (hidden) secondary structure states, *via* the posterior distribution $\Pr(c_i|S, \hat{T})$. These probabilities are calculated using the following formulae (in which all probabilities are conditional on $\hat{T}$, which is omitted for brevity):

$$\Pr(c_i|S_{i-1}) = \sum_{c_{i-1}} \Pr(c_i|S_{i-1}, c_{i-1})\Pr(c_{i-1}|S_{i-1})$$

$$= \sum_{c_{i-1}} \rho_{c_{i-1}c_i}\Pr(c_{i-1}|S_{i-1}) \qquad (8)$$

$$\Pr(c_i|S_i) = \frac{\Pr(s_i|S_{i-1}, c_i)\Pr(c_i|S_{i-1})}{\sum_{c_i} \Pr(s_i|S_{i-1}, c_i)\Pr(c_i|S_{i-1})}$$

$$= \frac{\Pr(s_i|c_i)\Pr(c_i|S_{i-1})}{\sum_{c_i} \Pr(s_i|c_i)\Pr(c_i\,/\,\mathrm{S}_{i-1})} \qquad (9)$$

$$\Pr(c_i|S) = \Pr(c_i|S_i)\sum_{c_{i+1}} \frac{\rho_{c_ic_{i+1}}\Pr(c_{i+1}|S)}{\Pr(c_{i+1}|S_i)} \qquad (10)$$

Equation (8) follows immediately from the law of total probability. Equation (9) is derived using Bayes' theorem, noticing that $\Pr(c_i|S_i) = \Pr(c_i|S_{i-1}, s_i)$ and that the independence of each alignment column s_i and the preceding columns S_{i-1} means that $\Pr(s_i|S_{i-1}, c_i) = \Pr(s_i|c_i)$. Equation (10) is derived by the following argument:

$$\Pr(c_i|S) = \sum_{c_{i+1}} \Pr(c_i, c_{i+1}|S)$$

$$= \sum_{c_{i+1}} \Pr(c_{i+1}|S)\Pr(c_i|c_{i+1}, S)$$

$$= \sum_{c_{i+1}} \Pr(c_{i+1}|S)\Pr(c_i|c_{i+1}, S_i)$$

$$= \sum_{c_{i+1}} \frac{\Pr(c_{i+1}|S)\Pr(c_i, c_{i+1}|S_i)}{\Pr(c_{i+1}|S_i)}$$

$$= \Pr(c_i|S_i)\sum_{c_{i+1}} \frac{\Pr(c_{i+1}|c_i)\Pr(c_{i+1}|S)}{\Pr(c_{i+1}|S_i)}\,.$$

Here, the first equality follows from the law of total probability, the second from the definition of conditional probability and the third from the conditional independence of c_i and s_j ($j > i$) given c_{i+1}. The last two equalities follow from the definition of conditional probability. Additional details are given by Churchill (1989).

The formulae are applied iteratively. Starting with the initial condition:

$$\Pr(c_1|S_0) = \sum_{c_0} \rho_{c_0c_1}\Pr(c_0|S_0) = \sum_{c_0} \rho_{c_0c_1}\Psi_{c_0} = \Psi_{c_1} \qquad (11)$$

Equations (8) and (9) define a recurrence relationship for $\Pr(c_i|S_i)$, the probabilities that column i is in each possible secondary structure state (c_i) given the data up to column i (S_i), *via* the intermediate value $\Pr(c_i|S_{i-1})$. They are used repeatedly with i increasing from 1 to N (the "forward pass") until $\Pr(c_N|S_N) \equiv \Pr(c_n|S)$ is calculated. Equation (10) is then used iteratively with i decreasing from $n - 1$ to 1 (the "backward pass") to calculate $\Pr(c_i|S)$ for each i, using the values of $\Pr(c_{i+1}|S_i)$ already calculated in the forward pass. The result is the posterior distribution $\Pr(c_i|S)$: for each site, the posterior probabilities that the site is in each of the secondary structure categories, given the data. Structure predictions are taken from these probabilities according to the rule that the predicted secondary structure category at a site is that with highest posterior probability.

Decision rules based on results other than the $\Pr(c_i|S)$ are possible, for example to choose the combination $C \equiv \{c_1, c_2, \ldots, c_N\}$ that makes the greatest contribution to the likelihood $\Pr(S|T) = \Sigma_C \Pr(S, C|T)$. We are currently investigating alternative decision rules.

Computer programs implementing the methods described here are available from N.G. (N.Goldman@gen.cam.ac.uk for FORTRAN code) and J.L.T. (thorne@stat.ascu.edu for C code) or by anonymous ftp from the directory pub/HMM at ftp.biochem.ucl.ac.uk.

Data

We illustrate our methods using xylanase sequences and close homologs. Xylanase is not represented in the 207 families of the DTJ-database. The xylanase A amino acid sequence for *Pseudomonas fluorescens* (SWISSPROT ID xyna_psefl, accession number P14768; we use the abbreviation PSEFL) was one of the target sequences in the recent Asilomar Prediction Challenge (Riddihough, 1994; Lattman, 1995; Moult *et al*., 1995). Its structure has subsequently been determined by X-ray diffraction to be a TIM-barrel [$(\beta/\alpha)_8$-barrel] (Harris *et al*., 1994; PDB entry 1XYS).

Database searches performed with the FASTA program (Pearson & Lipman, 1988) of the GCG package (Devereux *et al*., 1984) yielded six homologs with identities greater than 28% over a long part of the PSEFL sequence. We considered these sequences likely to have essentially the same secondary structure (Chothia & Lesk, 1986; Flores *et al*., 1993) and therefore to be suitable for alignment and analysis. The sequences were xylanase sequences from *Cellulomonas fimi* (gux_celfi, P07986; CELFI), *Bacillus stearothermophilus* (xyna_bacst, P40943; BACST), a thermophilic bacterium RT8.B4 (xyna_ther8, P40944; THER8) and a *Bacillus* species (strain C-125:

xyna_bacs5, P07528; BACS5), an endogluconase sequence from *Caldocellum saccharolyticum* (gunb_calsa, P10474; CALSA) and a celloxylanase sequence from *Clostridium stercorarium* (cexy_closr, P40942; CLOSR). Our results with these sequences are typical of a number of analyses that we have performed on target sequences from the Asilomar Prediction Challenge and other sources.

The sequences were aligned using the TREEALIGN program (Hein, 1990) and the resulting alignments were improved by hand. The final alignment we used, shown in Figure 2, contains 102 α-helix positions, 52 β-sheet positions and 155 loop positions.

Areas of the alignment not considered reliable were excluded from the analysis by replacing amino acids and gaps in uncertain sites by a missing data code in all but the PSEFL sequence. These missing data codes were treated in the manner described above for gaps. This procedure seems justified, as the pattern of amino acid replacements will be disrupted by alignment errors and because such unreliable regions may be those where the sequences are most likely to differ in secondary structure. Figure 2 indicates those parts of the xylanase alignment that were considered unreliable, and the true secondary structure as determined using DSSP. Sites where the multiple alignment procedure had introduced gaps into the PSEFL sequence of determined secondary structure were treated as being of unknown secondary structure, making no contribution to secondary structure prediction accuracy scores.

Results

HMM analyses

The ML estimate of the phylogenetic tree derived from our HMM for the xylanase data set is shown in Figure 3, and the corresponding secondary structure prediction is represented in Figure 4. The secondary structure prediction results are summarized in the all 7, HMM cell of Table 2. Mindful of the warning given by Jenny & Benner (1994), that there is more to secondary structure prediction than the overall proportion of positions correctly classified (Q_{total}), Table 2 records the proportion of positions of each secondary structure category correctly classified (Q_α, Q_β and Q_L). Table 2 records also the proportion of positions correctly predicted as not belonging to a particular secondary structure category among those positions that actually do not belong to that category (the "sensitivities", R_α, R_β and R_L of Ralph *et al*., 1987).

To illustrate the advantages of allowing dependencies introduced by common ancestry, we devised three experiments.

Effects of adding homologous sequences

First, to assess the effect on secondary structure prediction of including homologous sequences we

```
       1                                                50
         EEEEE             H HHHHHHHHH   EEEE           EE ? ?
PSEFL  FPIGVAVAAS GGNADIFTSS ARQNIVRAEF NQITAENIMK MSYMYSG-S-
CELFI  |||||||||| |||||||||| ||||IADSEF NLVVAENAMK WDATEPSQN-
CLOSR  FPIGAAI||| |||||||||| ||||LYKKHV NMLVAENAMK PASLQPTEG-
THER8  FKIGVAV||| |||||||||| ||||VIKRHF NSITPENEMK PESLQPYEG-
CALSA  FMIGVAI||| |||||||||| ||||MVLKHF NSITAENEMK PESLLAGQTS
BACST  FTIGAAV||| |||||||||| ||||MLKRHF NSIVAENVMK PISIQPEEG-
BACS5  FDIGAAV||| |||||||||| ||||ILKHHY NSLVAENAMK PESLQPREG-

       51                                              100
       ???EE  HHH HHHHHHHHH    EEEE                        HHH
PSEFL  ---NFSFTNS DRLVSWAAQN GQTVHGHALV WHPSYQLPNW ASDSNANFRQ
CELFI  ---SFSFGAG DRVASYAADT GKELYGHTLV WHS--QLPDW ||||||||||
CLOSR  ---NFQWADA DRIVQFAKEN GMELRFHTLV WHN--QTPTG ||||||||||
THER8  ---GFSFSIA DEYVDFCKKD NISLRGHTLV WHQ--QTPSW ||||||||||
CALSA  TGLSYRFSTA DAFVDFASTN KIGIRGHTLV WHN--QTPDW ||||||||||
BACST  ---KFNFEQA DRIVKFAKAN GMDIRFHTLV WHS--QVPQW ||||||||||
BACS5  ---EWNWEGA DKIVEFARKH NMELRFHTLV WHS--QVPEW ||||||||||

       101                                             150
       HHHHHHHHHH HH      EE EEEE                       HHHH
PSEFL  DFARHIDTVA AHFAGQVKSW DVVNEALFDS ADDPDGRGSA NGYRQSVFYR
CELFI  ||||HVTKVA DHFEGKVASW DVVNEAFADG DGPP------ --||||||||
CLOSR  ||||YIRAVV LRYKDDIKSW DVVNEVI||| ||||------ --||||||||
THER8  ||||HIQTVV GRYKGKVYAW DVVNEAI||| ||||------ --||||||||
CALSA  ||||YIYDVV GRYKGKVYAW DVVNEAI||| ||||------ --||||||||
BACST  ||||HIKTIV ERYKDDIKYW DVVNEVV||| ||||------ --||||||||
BACS5  ||||HIKTVV ERYKDDVTSW DVVNEVI||| ||||------ --||||||||

       151                                             200
       HH   HHHHH HHHHHH  ?     EEEEE         HHHH HHHHHHHHHH
PSEFL  QFGGPEYIDE AFRRAPRA-D PTAELYYNDF NTEENGAKTT ALVNLVQRLL
CELFI  ||LGNGYIET AFRAARAA-D PTAKLCINDY NVEGINAKSN SLYDLVKDFK
CLOSR  ||TGTEYIEV AFRATREAGG SDIKLYINDY NTDDP-VKRD ILYELVKNLL
THER8  ||LGPEYIEK AFIWAHEA-D PKAKLFYNDY STEDP-YKRE FIYKLIKNLK
CALSA  ||CGPEYIEK AFIWAHEA-D PNAKLFYNDY NTEIS-KKRD FIYNMVKNLK
BACST  ||AGIDYIKV AFQAARKYGG DNIKLYMNDY NTEVE-PKRT ALYNLVKQLK
BACS5  ||TGTDYIKV AFETARKYGG EEAKLYINDY NTEVP-SKRD DLYNLVKDLL

       201                                             250
       H      EEE E  EEE      HHHHHHHHH HHHH     E EEEEEEEEE
PSEFL  NNGVPIDGVG FQMHVMNDYP SIANIRQAMQ KIVALSPTLK IKITELDVRL
CELFI  ARGVPLDCVG FQSHLIVG-Q VPGDFRQNLQ RFADL--GVD VRITELDIRM
CLOSR  EKGVPIDGVG HQTHIDIYNP PVERIIESIK KFAGL--GLD NIITELDMSI
THER8  AKGVPVHGVG LQCHISLDWP DVSEIEETVK LFSRI-PGLE IHFTEIDISI
CALSA  SKGIPIHGIG MQCHINVNWP SVSEIENSIK LFSSI-PGIE IHITELDMSL
BACST  EEGVPIDGIG HQSHIQIGWP SEAEIEKTIN MFAAL--GLD NQITELDVSM
BACS5  EQGVPIDGVG HQSHIQIGWP SIEDTRASFE KFTSL--GLD NQVTELDMSL

       251                                             300
                               HHHHHHHH HHHHHHHHHH HH     EEE
PSEFL  NNPYDGNSSN DYTNRNDCAV SCAGLDRQKA RYKEIVQAYL EVVPPGRRGG
CELFI  RTPSDATK|| |||||||||| |||||||QAA DYKKVVQACM QVT---RCQG
CLOSR  ---Y-SWN|| |||||||||| |||||||QAK RYQELFDALK EN--KDIVSA
THER8  --------|| |||||||||| |||||||QAQ KLKAIFDVLK KY--RNVVTS
CALSA  ---Y-NYG|| |||||||||| |||||||QSQ KYKEIFTMLK KY--KNVVKS
BACST  ---Y-GWP|| |||||||||| |||||||QAA RYDRLFKLYE KL--SDKISN
BACS5  ---Y-GWP|| |||||||||| |||||||QAD RYDQLFELYE EL--AADISS

       301          315
       EEE              E
PSEFL  ITVWGIADPD SWLYT
CELFI  VTVWGITDKY SWVPD
CLOSR  VVFWGISDKY SWLNG
THER8  VTFWGLKDDY SWLRG
CALSA  VTFWGLKDDY SWLRS
BACST  VTFWGIADNH TWLDS
BACS5  VTFWGIADNH TWLDG
```

Figure 2. Amino acid sequence alignment of the seven xylanase homologs. Full species names and sequence accession numbers are given in the text. The first row of each block represents the true secondary structure. Secondary structure categories are represented by H for α-helix, E for β-sheet, a blank for loop, and ? for unknown. Dashes in the alignment represent gaps and vertical bars (|) represent areas treated as missing data because the alignment was deemed relatively uncertain.

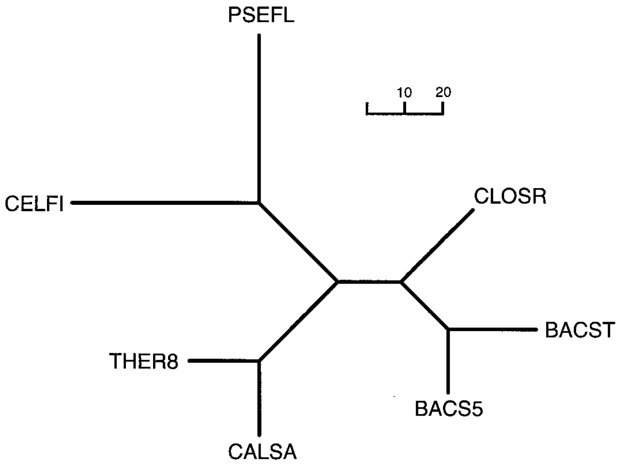

Figure 3. The ML estimate of phylogeny derived from our HMM method for the seven xylanase homologs. The scale-bar refers to the branch lengths, measured in units of expected number of replacements per 100 sites.

created three new data sets from the original alignment of Figure 2. The first (PSEFL only) consists of the PSEFL sequence alone; the second (close 3) comprises PSEFL and its two closest homologs, CELFI and CLOSR; the third (med 5) comprises the close 3 set plus THER8 and CALSA. Together with the original alignment (all 7), the progression PSEFL only–close 3–med 5–all 7 is intended to represent the gradual addition of information contained in patterns of amino acid replacement. The four data sets were analyzed with the HMM methods described above. In each case, the analysis included re-estimation of the phylogenetic tree. A summary of the results of secondary structure prediction is shown in Table 2. The contribution of the increasing number of sequences may be assessed by comparing results in the HMM column.

We find that overall prediction accuracy (Q_{total}) is higher when proteins related to PSEFL are included than when only PSEFL is considered. The improvement in Q_β, the prediction accuracy for β-sheets, is especially large. As more-distant relatives are incorporated into the analysis, overall prediction accuracy declines slightly, but it is always superior to prediction based on the PSEFL sequence alone. The decline may be due to the more distant relatives not sharing precisely the same secondary structure. Thus, the more distant homologs may introduce more noise than signal. This would be the case for all secondary structure prediction methods that use multiple sequence alignments.

In common with other studies (e.g. Lenstra *et al.*, 1977; Levin & Garnier, 1988; Rost & Sander, 1993), we find β-sheets most difficult to predict accurately. However, our β-sheet prediction scores (Q_β) continue to improve as even distant relatives of PSEFL are included in the analyses. This could indicate that β-sheet regions in the xylanase protein are well conserved across long evolutionary times, whereas α-helix regions are relatively less conserved. This would be in agreement with the observation that the β-barrel is the most highly conserved structural feature across TIM-barrel families (Pickett *et al.*, 1992; Sergeev & Lee, 1994).

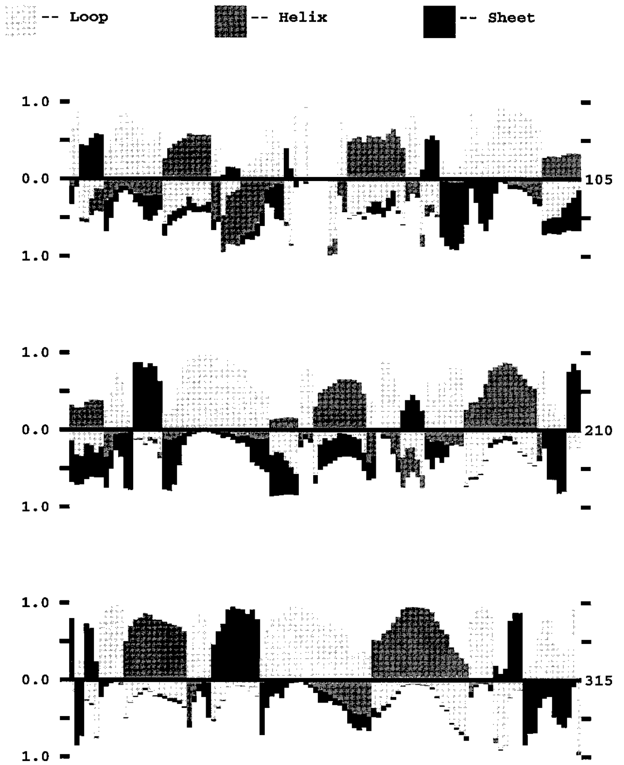

Figure 4. Graphical representation of our HMM secondary structure prediction for the *Pseudomonas fluorescens* xylanase sequence. The xylanase sequence alignment has 315 columns and the predictions at sites 1 to 105, sites 106 to 210 and sites 211 to 315 are indicated separately. For each site, the height of the column above the black horizontal line represents the posterior probability estimate for the true secondary structure category at that site. The posterior probability estimates of the other two secondary structure categories are indicated below the horizontal line. No estimate is shown for sites corresponding to insertions relative to the sequence of known structure. The shades that represent each of the three secondary structure categories are indicated on the Figure.

Effects of neglecting phylogeny

Second, we devised an experiment to assess the effect of considering phylogeny on secondary structure prediction. This is accomplished by comparison of the results described for the four data sets above with those obtained when the sequences within each data set are treated as though they were independent of one another. Specifically, we assume that residues from different sequences that are in a certain column of the alignment all belong to the same underlying secondary structure category but are otherwise independent. Our HMM method normally uses

Table 2. Accuracy of secondary structure prediction using the HMM method with and without consideration of the evolutionary tree and with duplicate copies of one sequence

	Analysis method							
Data set	HMM[a]		HMM (no tree)[b]		HMM (no tree) +2PSEFL[c]		HMM (no tree) + 9PSEFL[d]	
PSEFL only	**65.7**		**65.7**		**64.7**		**56.6**	
	72.5	*79.7*	72.5	*79.7*	82.4	*77.8*	61.8	*81.6*
	5.8	*95.7*	5.8	*95.7*	23.1	*86.8*	53.8	*75.5*
	81.3	*65.6*	81.3	*65.6*	67.1	*81.2*	54.2	*78.6*
close 3	**74.4**		**68.0**		**64.7**		**56.6**	
	82.4	*88.4*	80.4	*78.7*	81.4	*81.2*	58.8	*82.6*
	38.5	*94.2*	25.0	*92.6*	36.5	*83.3*	55.8	*74.7*
	81.3	*74.0*	74.2	*76.6*	63.2	*82.5*	55.5	*78.6*
med 5	**71.5**		**64.7**		**61.2**		**55.0**	
	73.5	*89.9*	66.7	*85.5*	63.7	*87.4*	53.9	*82.6*
	59.6	*89.5*	51.9	*81.7*	55.8	*75.1*	55.8	*73.2*
	74.2	*74.0*	67.7	*79.2*	61.3	*80.5*	55.5	*77.9*
all 7	**69.6**		**61.5**		**58.9**		**53.7**	
	66.7	*91.3*	59.8	*86.5*	58.8	*87.0*	52.9	*82.6*
	63.5	*84.0*	55.8	*77.8*	53.8	*75.1*	51.9	*72.8*
	73.5	*77.3*	64.5	*77.9*	60.6	*76.6*	54.8	*76.0*

As fully described in the text, variants of the HMM method are applied to four xylanase data sets. Each cell of the Table contains seven numbers: Q_{total} ; Q_{α} and R_{α}; Q_{β} and R_{β}; Q_{L} and R_{L}. Q_{total} is in bold; R_{α}, R_{β} and R_{L} are in italics. The PSEFL only, HMM and PSEFL only, HMM (no tree) cells are necessarily identical, as there is no evolutionary information in a single sequence.

[a] Unmodified HMM method.

[b] HMM method but with sequences treated as independent.

[c] As HMM (no tree) but with two duplicate copies of the PSEFL sequence added to each data set.

[d] As HMM (no tree) but with nine duplicate copies of the PSEFL sequence added to each data set.

the estimated phylogenetic tree, but this information can be excluded by re-running the HMM analysis with the sequences' phylogeny constrained to have infinitely long branch lengths (so the tree's topology is in fact irrelevant), in which case the sequences are indeed being treated as independent.

Secondary structure prediction summaries are shown in Table 2 under the heading HMM (no tree). Comparisons between the columns of Table 2 indicate the effect of using an evolutionary tree to model effects of common ancestry instead of treating sequences as independent. For the PSEFL only data set, the results are of course the same as there can be no evolutionary information in a single sequence. However, for the three other data sets we find there is a deterioration (6.8 to 9.7%) in Q_{total} when evolutionary information is ignored. Predictions of α-helices, β-sheets and loops deteriorate approximately equally when we ignore the evolutionary correlations in the data. All prediction accuracy criteria are affected approximately equally across the range of data sets, from those that include only the closer relatives of PSEFL to those that include the most divergent ones.

Effects of closely related homologs when phylogeny is neglected

The third experiment was designed to investigate further the effects of failing to model effects of common ancestry. We again applied this experiment to the four xylanase data sets, and performed the HMM (no tree) analysis as above on these alignments so that the sequences were treated as independent. To simulate the effects of having closely related sequences present in an alignment, we ran two related analyses for each data set: HMM (no tree) + 2PSEFL, in which two extra copies of the PSEFL sequence were added to the data sets, and HMM (no tree) + 9PSEFL, in which nine extra copies of PSEFL were added.

Adding exact copies of sequences to data sets analyzed by our (unmodified) HMM method has no effect on the results, as all equivalent sequences are (correctly) positioned in exactly the same place in the phylogenetic tree. Multiple exact copies therefore contribute no more information than a single copy. However, methods that do not carefully consider phylogenetic relationships are prone to give additional copies of one sequence undue influence on the secondary structure prediction. Although this is an artificial example of the consequences of ignoring common ancestry, we believe that it may not be too different from the situation in real data sets where inappropriate weight could be given to relatively similar sequences.

The secondary structure prediction results of these analyses are summarized in the last two columns of Table 2. Comparisons between these columns and the HMM (no tree) column indicate a further effect of biasing analyses by failing to allow for phylogenetic relationships of closely related sequences. For each data set, the trend is for the overall accuracy (Q_{total}) to decline as more duplicate sequences are added. This can be attributed in particular to a fall in Q_{L}. The trend is also for prediction accuracy to decline as more-distantly related sequences are included (movement down the appropriate columns of Table 2). This decline is attributable to deterioration in Q_{α} and, to a lesser extent, Q_{L}.

These effects can be explained as follows. As more copies of the PSEFL sequence are added with no account taken of their relationships to one another, each sequence position appears increasingly conserved as more sequences share the same residue. This (mis)information leads to the (mis)interpretation that the secondary structure for which that residue is most probable (i.e. the k that maximizes $\pi_i^{(k)}$, where the repeated residue is i; see Theory) is overwhelmingly the most likely. The Markov model of organization of structure along sequences becomes irrelevant, and secondary structure prediction depends only on the PSEFL sequence and the secondary structures that make each of its residues (independently) most likely.

Other analyses

The PHD program (Rost & Sander, 1993, 1994; Rost *et al*., 1994) is among the best automated predictors of secondary structure but it does not explicitly consider phylogenetic relationships. Therefore, we use the PHD program to further illustrate the effects of failing to model common ancestry of related sequences. The four xylanase data sets were analyzed using PHD, and re-analyzed with two or nine extra copies of the PSEFL sequence added (PHD + 2PSEFL and PHD + 9PSEFL, respectively). We did not use the optional PHD facilities to search databases for homologs and align sequences; in addition, since PHD does not accept ‘‘unknown residue’’ codes in sequences (analogous to our use of missing data codes for unreliable alignment sites), we recoded the uncertain alignment areas (Figure 2) as gaps. This procedure was designed to make the assumptions and data of our HMM method and PHD as close as possible; however, we could not match them completely. Importantly, PHD allows the positions of gaps in an alignment to affect secondary structure predictions. This is desirable because gaps tend to be found in loop regions (Benner & Gerloff, 1991; Thornton *et al*., 1991). Since it is expected that the uncertain areas of our alignment are in loop regions, and because loop regions may be less constrained by selection, PHD will likely have a qualitative advantage over our HMM method for these areas.

The secondary structure prediction results of the PHD, PHD + 2PSEFL and PHD + 9PSEFL analyses are summarized in Table 3. Comparisons between columns again indicate a biasing effect that the appropriate representation of common ancestry would eradicate. Results of the three analyses are identical for the PSEFL only data set, all sequence comparisons being between identical sequences. However, for all the other data sets it is clear that the duplicate sequences, although they can contain no new information, are having an effect on secondary structure prediction. This is evident in the prediction scores Q and R, which vary along rows of Table 3.

Table 3 also records the results of PHD’s subset predictions, which indicate residues where the PHD method is particularly confident about its predictions. Although there is no obvious trend in the overall prediction accuracy (Q_{total}) for those predictions that PHD makes with this higher level of confidence, we note that the number of residues that are predicted with this confidence tends to increase as exact copies of PSEFL are added to the data set. The PHD program gives these sequences undue influence; the copies are being treated by PHD as if they contribute relevant additional information. In contrast, our method completely discounts the duplicates because they are perfectly correlated with the original PSEFL sequence. The PHD program lacks an explicit consideration of phylogenetic relationships, as do many other secondary structure prediction methods, and therefore appropriate discounting of information is not performed.

Table 3. Performance of secondary structure prediction using the PHD method and the PHD method with additional copies of the PSEFL sequence

	Analysis method					
Data set	PHD		PHD+ 2PSEFL		PHD+ 9PSEFL	
PSEFL only	**71.8**		**71.8**		**71.8**	
	82.4	*84.1*	82.4	*84.1*	82.4	*84.1*
	61.5	*88.3*	61.5	*88.3*	61.5	*88.3*
	68.4	*84.4*	68.4	*84.4*	68.4	*84.4*
	(78.7, 0.73)		(78.7, 0.73)		(78.7, 0.73)	
close 3	**77.4**		**76.1**		**76.1**	
	89.2	*85.0*	86.3	*85.5*	86.3	*85.0*
	55.8	*93.4*	57.7	*91.8*	57.7	*91.4*
	76.8	*85.7*	75.5	*85.1*	75.5	*85.1*
	(82.7, 0.72)		(83.2, 0.75)		(83.0, 0.74)	
med 5	**79.9**		**78.6**		**78.3**	
	89.2	*89.9*	85.3	*89.4*	85.3	*90.8*
	63.5	*91.4*	63.5	*91.4*	65.4	*90.7*
	79.4	*87.7*	79.4	*85.7*	78.1	*84.4*
	(89.9, 0.66)		(89.5, 0.67)		(87.7, 0.67)	
all 7	**79.6**		**78.6**		**77.0**	
	90.2	*88.4*	88.2	*88.4*	86.3	*88.9*
	63.5	*91.4*	63.5	*91.4*	59.6	*90.3*
	78.1	*89.0*	77.4	*87.0*	76.8	*85.1*
	(89.1, 0.67)		(90.0, 0.67)		(88.1, 0.69)	

The PHD-based methods are applied to four xylanase data sets, as described in the text. The top seven values in each cell of the Table are as in Table 2; beneath these, in parentheses, are Q_{total} for PHD’s subset of high accuracy followed by the fraction of residues in this subset.

Discussion

It has been suggested that incorporating evolutionary information into protein secondary structure prediction methods can improve their accuracy by approximately 7.5% (Rost & Sander, 1994). Comparisons within appropriate columns of Tables 2 and 3 corroborate this suggestion. The results shown in the first two columns of Table 2 illustrate how our HMM separates the contributions to secondary structure prediction of multiple homologous sequences and their evolutionary relatedness. These results show that the explicit modelling of evolutionary relatedness can add to the accuracy of predictions of secondary structure, beyond the improvement generated simply by using multiple sequences.

The results shown in the last two columns of Table 2 and in Table 3 demonstrate further deleterious effects of failing to model evolutionary relationships in comparative sequence analyses. Although the addition of closely related and therefore very similar sequences should add little or no information, these results indicate the actual consequences for our HMM method, when modified to neglect phylogeny, and for PHD. The specific reasons for the results obtained are easily explained for our method (above). The results from PHD are harder to explain precisely, as its neural network jury architecture acts rather like a ‘‘black box’’. Nevertheless, the effects of neglecting phylogeny are qualitatively the same; that is, a decrease in ability to predict secondary structure accurately. We expect this finding to apply to any comparative sequence analysis that ignores the evolutionary relationships between the sequences it analyzes. The appropriate way to assess the contributions of multiple homologous sequences is *via* the phylogeny but no previous method employs phylogenetic information directly. Because of this, we believe that further improvements will be obtained by modifying the best available secondary structure prediction methods so that evolutionary relationships among sequences are explicitly considered.

Although one important advantage of our HMM method is its treatment of the correlations among homologous sequences, our method has other desirable properties. Sites in different secondary structure categories may tend to experience both different rates and different patterns of amino acid replacement. Other secondary structure prediction methods do not consider evolution and are therefore apt to confound the information from the rates and patterns of amino acid replacements.

We acknowledge that our HMM method is not at this stage a serious contender amongst the best secondary structure prediction methods. The best methods consider features of data that our method ignores. We certainly do not claim any superiority over the highly successful PHD method, which we have used here simply for illustrative purposes. We chose the PHD program because it is one of the most widely used methods, and is available freely *via* electronic mail (Rost *et al*., 1994). For these reasons, we have deliberately avoided any comparisons of secondary structure prediction accuracy between our HMM method and PHD or other prediction methods. If such a comparison is required, the single ‘‘fair’’ comparison available here is between the all 7, HMM cell of Table 2 and the all 7, PHD cell of Table 3. We prefer to stress the results of the experiments described above, especially the second, as indicators of the improvements that might be possible in many other comparative sequence analysis methods if phylogenetic relationships were to be explicitly modelled.

We admit that our model is quite simplistic. It is certainly not true that secondary structure lengths follow geometric distributions, as our Markov chain model implies. The three secondary structure categories we have used are not all the possible natural ones and the model does not utilize information in alignment gaps. The amino acid replacement model would undoubtedly be improved by consideration of the three-dimensional structure of proteins (e.g. distinguishing interior and surface sites; see Koshi & Goldstein, 1995), and this modification would surely permit more information on secondary structure to be extracted (Thompson & Goldstein, 1996a,b). Such improvements are under consideration. However, our model already appears to be better than current evolutionary models used in phylogenetic analyses for some proteins, and so may be expected to give better estimates of phylogenetic relationships than other current methods (Thorne *et al*., 1996).

Improvements to our HMM are expected to lead to better secondary structure predictions, as might some form of ‘‘filtering’’ to modify ‘‘nonsense’’ predictions (e.g. single residue α-helices or β-sheets) that are currently permitted. Another improvement could involve preliminary analysis to classify proteins as all-α, all-β, α/β, etc. followed by secondary structure prediction with HMMs optimized for these classes of proteins (cf. Garnier *et al*., 1978; Rost & Sander, 1993; Stultz *et al*., 1993).

Whether such modifications ever make our method a reliable secondary structure prediction tool, we stress that the proper treatment of phylogenetic correlations within sequence alignments can improve methods that operate on evolutionarily related molecular sequences. In addition, explicit modelling permits insights beyond secondary structure prediction and may generate the understanding to improve methods further in a way that other approaches, for example neural networks, cannot.

Acknowledgements

We thank Mark Bradburn and Burkhard Rost for assistance with submitting alignments to the PHD e-mail server, and Richard Goldstein for providing copies of papers in press. N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research. J.L.T.'s research was supported by NIH grant R01 GM 49654 and the Alfred P. Sloan Foundation. D.T.J. is supported by a Royal Society University Research Fellowship.

References

Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **207**, 647–653.

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

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enz. Reg.* **31**, 121–181.

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324.

Bleasby, A. J. & Wootton, J. C. (1990). Construction of validated, non-redundant composite protein sequence databases. *Protein Eng.* **3**, 153–159.

Chothia, C. & Lesk, A. M. (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.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the α-subunit of tryptophan synthase. *Proteins: Struct. Funct. Genet.* **2**, 118–129.

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

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

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.

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. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Gen. Res.* **59**, 139–147.

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

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

Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Gerstein, M., Sonnhammer, E. L. L. & Chothia, C. (1994). Volume changes in protein evolution. *J. Mol. Biol.* **236**, 1067–1078.

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

Harris, G. W., Jenkins, J. A., Connerton, I., Cummings, N., Lo Leggio, L., Scott, M., Hazlewood, G. P., Laurie, J. I., Gilbert, H. J. & Pickersgill, R. W. (1994). Structure of the catalytic core of the family F xylanase from *Pseudomonas fluorescens* and identification of the xylopentaose-binding sites. *Structure*, **2**, 1107–1116.

Harvey, P. H. & Pagel, M. D. (1991). *The Comparative Method in Evolutionary Biology*, Oxford University Press, Oxford.

Harvey, P. H. & Purvis, A. (1991). Comparative methods for explaining adaptations. *Nature*, **351**, 619–624.

Hein, J. (1990). A unified approach to alignment and phylogenies. *Methods Enzymol.* **183**, 626–645.

Jenny, T. F. & Benner, S. A. (1994). Evaluating predictions of secondary structure in proteins. *Biochem. Biophys. Res. Commun.* **200**, 149–155.

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

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, H. N., ed.), pp. 21–132, Academic Press, New York.

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

Kishino, H., Miyata, T. & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**, 151–160.

Koshi, J. M. & Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein Eng.* **8**, 641–645.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology—applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Lattman, E. E. (1995). Protein structure prediction – a special issue. *Proteins: Struct. Funct. Genet.* **23**, R1.

Lenstra, J. A., Hofsteenge, J. & Beintema, J. J. (1977). Invariant features of the structure of pancreatic ribonuclease: a test of different predictive models. *J. Mol. Biol.* **109**, 185–193.

Levin, J. A. & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta*, **955**, 283–295.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995).

A large-scale experiment to assess protein structure prediction methods. *Proteins: Struct. Funct. Genet.* **23**, R2–R4.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Pickett, S. D., Saqi, M. A. S. & Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction—discrimination of the $(\beta/\alpha)_8$-barrel fold. *J. Mol. Biol.* **228**, 170–187.

Ralph, W. W., Webster, T. & Smith, T. F. (1987). A modified Chou and Fasman protein structure algorithm. *CABIOS*, **3**, 211–216.

Riddihough, G. (1994). Challenging the predictors. *Nature Struct. Biol.* **1**, 265–266.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.

Rost, B., Sander, C. & Schneider, R. (1994). PHD – an automatic mail server for protein secondary structure prediction. *CABIOS*, **10**, 53–60.

Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11–15.

Sergeev, Y. & Lee, B. (1994). Alignment of β-barrels in $(\beta/\alpha)_8$ proteins using hydrogen-bonding pattern. *J. Mol. Biol.* **244**, 168–182.

Stultz, C. M., White, J. V. & Smith, T. F. (1993). Structural analysis based on state-space modeling. *Protein Sci.* **2**, 305–314.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). Phylogenetic Inference. In *Molecular Systematics* (Hillis, D. M., Moritz, C. & Mable, B. K., eds), 2nd edit., pp. 407–514, Sinauer Associates, Sunderland, MA.

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

Thompson, M. J. & Goldstein, R. A. (1996a). Constructing amino acid residue substitution classes maximally indicative of local protein structure. *Proteins: Struct. Funct. Genet.* **25**, 28–37.

Thompson, M. J. & Goldstein, R. A. (1996b). Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins: Struct. Funct. Genet.* **25**, 38–47.

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

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1991). Prediction of progress at last. *Nature*, **354**, 105–106.

Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. P. & Blundell, T. L. (1993). Fragment ranking in modeling of protein structure: conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **229**, 194–220.

Wako, H. & Blundell, T. L. (1994a). 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.

Wako, H. & Blundell, T. L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol.* **238**, 693–708.

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

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

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by F. E. Cohen

(Received 9 May 1996; received in revised form 5 August 1996; accepted 12 August 1996)