JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 233–248

Bayesian Segmentation of Protein Secondary Structure

SCOTT C. SCHMIDLER,[1,2,3] JUN S. LIU,[2] and DOUGLAS L. BRUTLAG[3]

ABSTRACT

We present a novel method for predicting the secondary structure of a protein from its amino acid sequence. Most existing methods predict each position in turn based on a local window of residues, sliding this window along the length of the sequence. In contrast, we develop a probabilistic model of protein sequence/structure relationships in terms of structural *segments*, and formulate secondary structure prediction as a general Bayesian inference problem. A distinctive feature of our approach is the ability to develop explicit probabilistic models for α-helices, β-strands, and other classes of secondary structure, incorporating experimentally and empirically observed aspects of protein structure such as helical capping signals, side chain correlations, and segment length distributions. Our model is Markovian in the *segments*, permitting efficient exact calculation of the posterior probability distribution over all possible segmentations of the sequence using dynamic programming. The optimal segmentation is computed and compared to a predictor based on marginal posterior modes, and the latter is shown to provide significant improvement in predictive accuracy. The marginalization procedure provides exact secondary structure probabilities at each sequence position, which are shown to be reliable estimates of prediction uncertainty. We apply this model to a database of 452 nonhomologous structures, achieving accuracies as high as the best currently available methods. We conclude by discussing an extension of this framework to model nonlocal interactions in protein structures, providing a possible direction for future improvements in secondary structure prediction accuracy.

Key words: protein secondary structure prediction, Bayesian methods, probabilistic modeling.

1. INTRODUCTION

PREDICTION OF THE SECONDARY STRUCTURE of a protein from its amino acid sequence remains an important and difficult task. Not only can successful predictions provide a starting point for direct tertiary structure modeling (Friesner and Gunn, 1996; Jones *et al.*, 1994; Monge *et al.*, 1994; Rost *et al.*, 1996), but they can also significantly improve sequence analysis and sequence-structure threading (Fischer and Eisenberg, 1996; Russell *et al.*, 1996) for aiding in structure and function determination. However, despite considerable progress in secondary structure prediction over the last decade (see Barton (1995)

[1] Section on Medical Informatics, Stanford University School of Medicine, Stanford, CA 94305.
[2] Department of Statistics, Stanford University, Stanford, CA 94305.
[3] Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305.

for a recent survey), the current best methods reach accuracies of about 75% when multiple homologous sequences are available (Frishman and Argos, 1997), and 71% for single sequence predictions (Salamov and Solovyev, 1997). New methods which more accurately reflect features of protein structure folding and stabilization may be necessary to advance prediction beyond these levels.

Since early attempts to predict secondary structure (Garnier *et al.*, 1978), most efforts have focused on development of mappings from a local window of residues in the sequence to the structural state of the central residue in the window, and a large number of methods for estimating such mappings have been developed. Early approaches scored individual amino acids by frequency of occurrence in each structural state, combining them in ways corresponding to conditional independence models (Chou and Fasman, 1974; Garnier *et al.*, 1978). Improvements in accuracy were achieved by methods considering correlations among positions within the window, either implicitly using semi- and non-parametric statistical models such as neural networks (Holley and Karplus, 1989; Qian and Sejnowski, 1988; Stolorz *et al.*, 1992) and nearest-neighbor classifiers (Yi and Lander, 1993; Zhang *et al.*, 1992), or explicitly (Garnier *et al.*, 1996; Munson *et al.*, 1994; Riis and Krogh, 1996). Further improvements were demonstrated by the inclusion of evolutionary information via multiple alignments of homologous sequences (Di Francesco *et al.*, 1996; Rost and Sander, 1993a; Rost and Sander, 1994; Salamov and Solovyev, 1995), although the relative contribution of such information has been debated (Benner, 1995; Frishman and Argos, 1996). Interestingly, most recent improvements in accuracy have come from methods which are capable of considering *nonlocal* interactions in the sequence which occur outside a fixed length window (Frishman and Argos, 1996, 1997; Salamov and Solovyev, 1997). Here we take a *model-based* approach, formulating secondary structure prediction as a general Bayesian inference problem. Such an approach avoids many of the problems associated with window-based predictions, such as the need for post-prediction "filtering" (Frishman and Argos, 1996; Rost and Sander, 1993b) and provides a general framework for incorporation of the growing body of scientific knowledge about protein structure into the prediction process.

2. METHODS

We begin by choosing a representation of sequence/structure relationships in proteins which is based on *segments* of secondary structure. We parameterize this model in a convenient fashion by representing the segment positions and structural types. We denote segment locations by the position of the last residue in the segment, following Auger and Lawrence (1989) and Liu and Lawrence (1996). Because segments are required to be contiguous, this parameterization uniquely identifies a set of segment locations for a given sequence.

Let $R = (R_1, R_2, \ldots R_n)$ be a sequence of n amino acid residues, $S = \{i : \mathit{Struct}(R_i) \neq \mathit{Struct}(R_{i+1})\}$ be a sequence of m positions denoting the end of each individual structural segment (so that $S_m = n$), and $T = (T_1, T_2, \ldots, T_m)$ be the sequence of secondary structural types for each respective segment. An example is given in Figure 1. We will concern ourselves with the 3-state problem, where $\forall i T_i \in \{H, E, L\}$, although generalizations may be desirable. Together m, S and T completely determine a secondary structure assignment for a given amino acid sequence. In the case of secondary structure prediction, the quantities of interest are thus the values of m, $S = (S_1, S_2, \ldots, S_m)$ and $T = (T_1, T_2, \ldots, T_m)$ corresponding to the known amino acid sequence $R = (R_1, R_2, \ldots R_n)$, i.e., the locations and types of the secondary structural segments. The problem is to infer the values of (m, S, T) given a residue sequence R.

We take a Bayesian approach to the assignment of these parameter values, by defining a joint probability distribution $P(R, m, S, T)$ for an amino acid sequence and its secondary structure assignment. We then compute the conditional or posterior probability distribution over structural assignments given a new sequence $P(m, S, T \mid R)$ via Bayesian inference, and predict those secondary structure assignments (m, S, T) which maximize this posterior distribution. In Section 2.1, we define a general segment-based joint probability model which lends itself to efficient exact calculation of the posterior. Section 2.2 provides

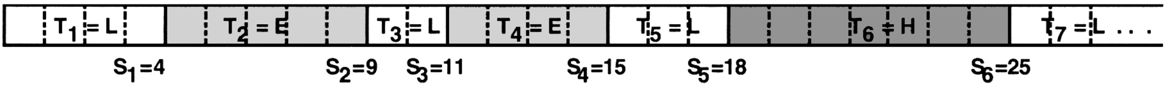


FIG. 1. Representation of the secondary structure of a protein sequence in terms of structural *segments*. The parameters shown represent the segment types T = (L,E,L,E,L,H,L, . . .) and endpoints S = (4,9,11,15,18,25, . . .). The associated structure assignment is LLLLEEEEELLEEEELLLHHHHHHHLLL . . . .

specific models for α-helices, β-strands, and loops, and shows how such models can be used to capture key aspects of protein structure formation observed in experimental settings and database analyses. Section 2.3 describes an algorithm for computation of quantities of interest under the posterior distribution.

2.1. Basic model

A key aspect of our approach is the choice of a joint probability model $P(R, m, S, T)$ which is decomposable into individual segment terms. In other words, the joint distribution may be factored by conditional independence of inter-segment residues, so that the sequence likelihood can be written as a product of segment likelihoods:

$$P(R \mid m, S, T) = \prod_{j=1}^{m} P(R_{[S_{j=1}+1:S_j]} \mid S, T) \tag{1}$$

where the jth term on the right-hand side of (1) is the likelihood of the subsequence of R beginning at position $S_{j-1} + 1$ and ending at position S_j, in other words, the amino acids in segment j. The exact form of this segment likelihood is structure-dependent, and the specification of this form for each structural type amounts to developing a probabilistic *model* of the given type of segment. The particular models used in this paper are developed in Section 2.2, but some general comments are appropriate here. First, note that this model does not assume conditional independence of int*ra*-segment residues; in fact, as described in Section 2.2, an explicit goal of our approach is to choose a form which allows us to model correlation among positions within a segment. Moreover, the terms $P(R_{[S_{j-1}+1:S_j]} \mid S, T)$for individual segments can take on arbitrary form and may depend on general properties of a segment (such as hydrophobic moment or helix dipole) beyond properties of individual residues.

Given (1), we need only provide the prior distribution $P(m, S, T)$ to completely specify the joint distribution $P(R, m, S, T)$. A computationally convenient choice is to factor $P(S, T \mid m)$ as a Markov process:

$$P(m, S, T) = P(m) \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j) \tag{2}$$

where each segment type depends only on its nearest neighbors, and the conditioning of S_j on (S_{j-1}, T_j) allows explicit modeling of the differing length distributions of each segment type observed in the Protein Data Bank (Bernstein *et al.*, 1977), as shown in Figure 2. Here we take $P(m)$ to be improper uniform; more informative priors on m are possible, but have little impact (Schmidler, 2000). By the choice of (2), our model becomes closely related to the class of *hidden semi-Markov* or *semi-Markov source* models discussed in Levinson (1986), Rabiner (1989), and Russell and Moore (1985) for applications in speech recognition. In the speech recognition literature, however, observations during a given state occupancy are typically modeled as *iid* (independent and identically distributed). As described in Section 2.2, the ability to model both nonindependence and nonidentity of distributions is the major motivation for our segment-based approach. We note also that a model very similar to that given in (1) and (2) has been developed independently by Burge and Karlin (1997) and applied to gene parsing in eukaryotic DNA with great success.

2.2. Probabilistic models for protein structure

Our goal here is to choose a specific form of the segment likelihood $P(R_{[S_{j-1}+1:S_j]} \mid S, T)$ which captures core aspects of protein secondary structure formation: amino acid propensities, hydrophobicity patterns, side chain interactions, and helical capping signals. In other words, we wish to develop probabilistic *models* for protein structural segments. For example, the function $P(R_{[i:j]} \mid i, j, H)$ provides the likelihood of the subsequence $R_{[i:j]}$ under the assumption that a helix begins at position i and ends at position j. Given such a segment likelihood for each structural class (H, E, L), computing the likelihood of a sequence under any given structural assignment is trivially done by evaluating the product of (1) and (2). Here we provide the exact forms for these segment likelihoods used in this paper.

Helix model. The presence of correlated side chain mutations in α-helices has been well studied, deriving from both environmental constraints such as hydropathy (Eisenberg *et al.*, 1984) and from stabilizing side chain interactions (Klingler and Brutlag, 1994). These correlations in nonadjacent sequence positions

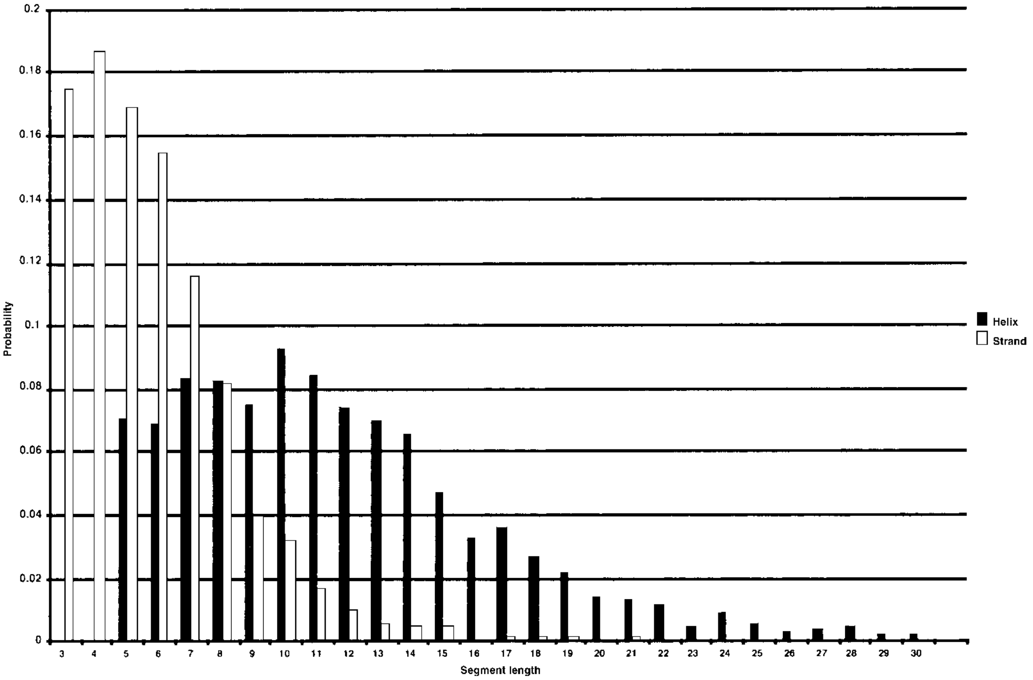


FIG. 2. Empirical length distribution of observed structural segments for α-helices (dark) and β-strands (light). Distributions are calculated from the structural database described in Section 3.

are induced by their spatial proximity in the folded protein molecule and hence provide an important source of information about the underlying structure. Figure 3 shows an example of an amphipathic α-helix which exhibits periodicity in sequence hydrophobicity. Because of the differing rates of rotation in helices and strands, this side chain periodicity can be an important clue for identifying the underlying backbone conformation.

Another important source of information for identifying α-helical segments in protein sequences is the existence of *helical capping* signals, the preference for particular amino acids at the N- and C-terminal ends which terminate helices through side chain-backbone hydrogen bonds or hydrophobic interactions. Such signals have been well characterized experimentally in terms of their stabilizing effect in helical peptides (Doig and Baldwin, 1995; Presta and Rose, 1988; Richardson and Richardson, 1988) (see Aurora and Rose (1998) for a review), as well as empirically through observed correlations (Klingler and Brutlag, 1994). This capping effect results in amino acid distributions at end-segment positions which differ significantly from those of internal positions. Figure 4 shows some of these informative distributions.

Our goal is to develop a helical segment model which captures such position-specific preferences and probabilistic dependence of int*ra*-segment residues, in addition to standard amino acid propensities. The model must also account for helices of various lengths. In this paper we use the following form of this distribution:

$$
\begin{aligned}
P(R_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, H) = & \prod_{i=S_{j-1}+1}^{S_{j-1}+\ell_N^H} P^H_{N_{i-S_{j-1}}}(R_i \mid R_{[S_{j-1}+1:i-1]}) \\
& \times \prod_{i=S_{j-1}+\ell_N^H+1}^{S_j-\ell_C^H} P_I^H(R_i \mid R_{[S_{j-1}+1:i-1]}) \\
& \times \prod_{i=S_j-\ell_C^H+1}^{S_j} P^H_{C_{S_j-i+1}}(R_i \mid R_{[S_{j-1}+1:i-1]}).
\end{aligned}
\qquad (3)
$$

Here ℓ_N^H indicates the length of the helix N-cap model, N_i, C_i indicate the ith position from the N- and C-termini respectively; and I indicates an internal (noncap) position. Figure 5 shows graphically how this

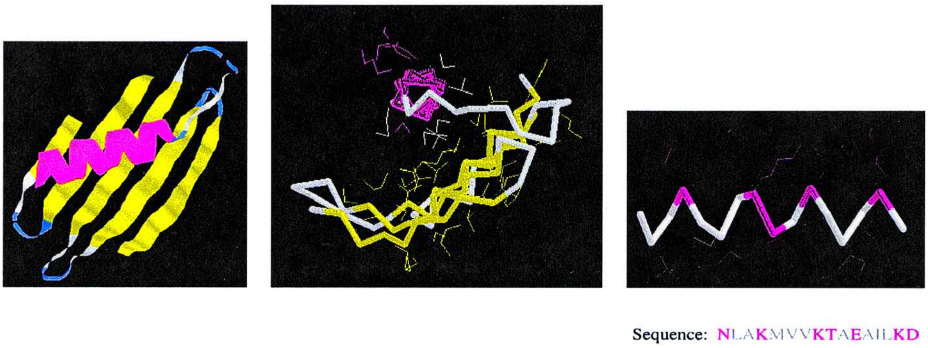


FIG. 3. Amphipathic helix from β-lactamase (4blm) showing hydrophobic side chain periodicity induced in sequence. The first image shows the helix in its native environment in the folded structure. The second image shows that this amphipathic environment induces a distinct preference for hydrophobic side chains on the buried surface of the helix (hydrophobic side chains are shown in white). The third image demonstrates how this preference in combination with the rotation of the helix induces a distinct periodicity in the amino acid sequence.

model is applied to the particular amino acid subsequence of the helix in Figure 3: the first product term in (3) models the distribution of amino acids at each of the first ℓ_N^H N-terminal positions (N-cap, N1, N2, N3, ...), and similarly the last term for the C-terminal positions (..., C3, C2, C1, C-cap), while the middle term models all internal positions as identically distributed but dependent.

In choosing the length of the helix cap models ℓ_N^H, ℓ_C^H, we considered caps of up to 4 positions at each end of the segment. The first 4 such positions at each terminus in α-helices are of particular interest due to their inability to form intra-helical hydrogen bonds, their propensity for acidic/basic side chains, and stabilization effects of the helical dipole moment. Figure 4A shows distributions for these positions in α-helices. As described in Section 3, we use secondary structure assignments provided by DSSP (Kabsch and Sander, 1983) which do not include the first and last hydrogen bonded residues in a helix. Hence the N-cap and C-cap positions are not typically included by DSSP. (To correct for this, we allow the segment transition term in (2) to depend on the last residue of the previous segment.) Nevertheless, Figure 4A displays previously observed patterns such as the prevalence of Pro at position N1 and Glu and Asp at position N2, while Figure 4b shows the expected prevalence of Ala and various hydrophobic residues at internal positions of helices.

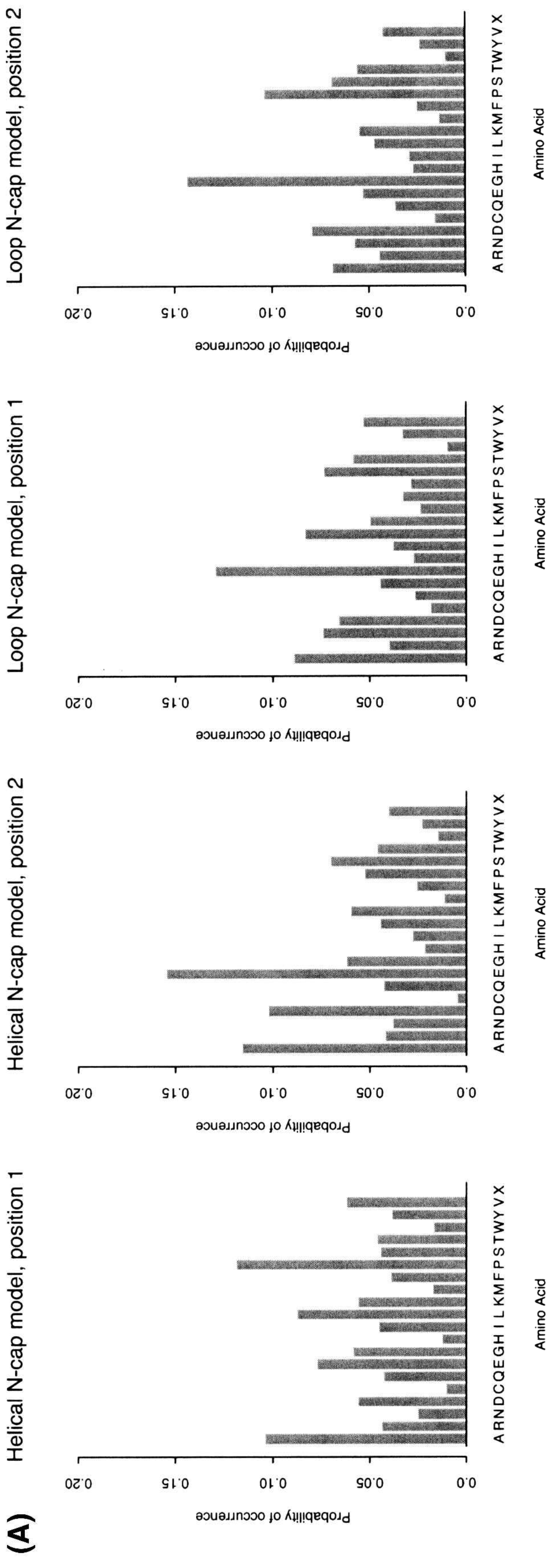
Table 1 shows the statistical deviance between the amino acid distribution at each end-segment position and the amino acid distribution at internal positions, calculated using the data set described in Section 3. The strongest signal appears in the first two positions of the helical N-terminus (N1 and N2), while β-strands and loops show little change in these positions. The positions we included in each structural model for predictive purposes are highlighted (so that $\ell_N^H = 4$, $\ell_C^H = 1$), capturing the positions that are significantly different. It should be noted that such information is inherently difficult to include in window-based prediction methods, which must scan a residue across each position in the window in turn.

Equation (3) provides everything except the exact intra-segment residue dependencies in the model. For α-helices, these are given by:

$$P_i^H(R_i \mid R_{[j:i-1]}) = P_i^H(R_i \mid h_i)P_i^H(h_i \mid h_{i-2}, h_{i-3}, h_{i-4}) \tag{4}$$

where $h_i \in \{\textit{hydrophobic, neutral, hydrophilic}\}$ indicates the hydrophobicity class of residue R_i assigned by Klingler and Brutlag (1994). In other words, dependency between positions is modeled using a reduced alphabet in order to avoid combinatorial explosion of parameters. Figure 6 provides a graphical model representation (Whittaker, 1990) of the dependency structure given by (4). This form of the distribution allows us to explicitly capture the previously described intra-segment residue correlations corresponding to the periodicity of an α-helix, by conditioning the probability of a particular residue on the $i-4$, $i-3$, and $i-2$ residues. Internal positions are therefore modeled as identically distributed, but dependent. We note that Stultz *et al.* (1993) also provide a model for amphipathicty in α-helices in their development of structured hidden Markov models for particular tertiary folds.

β-Strand and Loop models. The general form of (3) is convenient for modeling variable-length segments, and we retain such a form for β-strand and loop segments. However, the utility of distinguishing


(A)
Helical N-cap model, position 1
Helical N-cap model, position 2
Loop N-cap model, position 1
Loop N-cap model, position 2
Probability of occurrence
0.0
0.05
0.10
0.15
0.20
A R N D C Q E G H I L K M F P S T W Y V X
Amino Acid

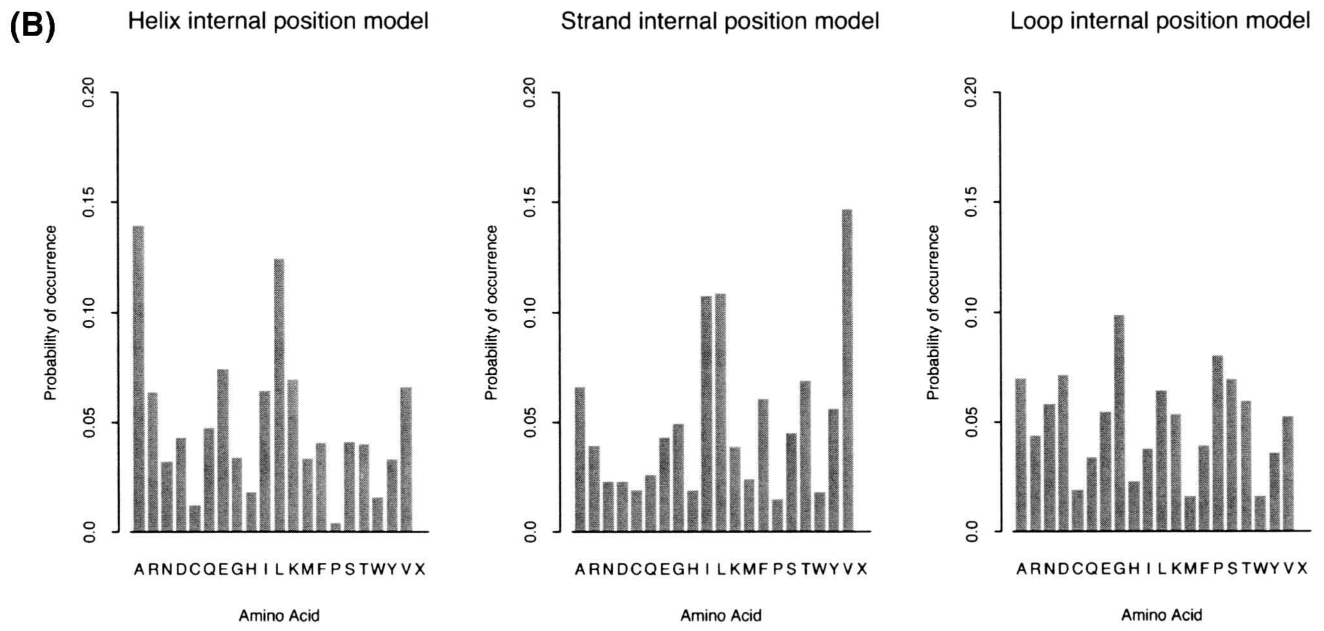
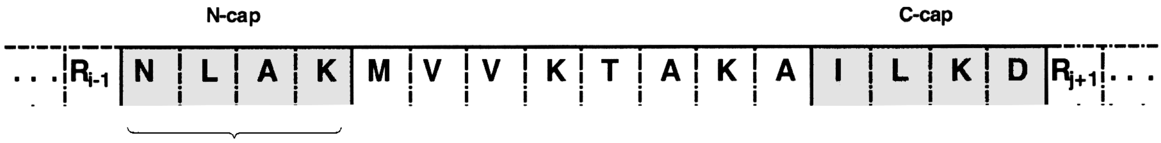


FIG. 4. Amino acid distributions for (**A**) the first and second N-terminal positions (N1 and N2) in α-helices (panels 1 and 2) and loops/coils (panels 3 and 4), and (**B**) internal positions in helices (panel 1), strands (panel 2), and loops/coils (panel 3). The distributions display well-known preferences for particular amino acids based on structural type and position as described in Section 2.2. Distributions are calculated from the structural database described in Section 3.



$$P^H_{N_0}(N)P^H_{N_1}(L \mid N)P^H_{N_2}(A \mid N,L)P^H_{N_3}(K \mid N,L,A)$$

$$\times \quad P^H_I(M \mid N,\ldots,K)P^H_I(V \mid N,\ldots,M)\ldots P^H_I(A \mid N,\ldots,K)$$

$$\times \quad P^H_{C_3}(I \mid N,\ldots,A)P^H_{C_2}(L \mid N,\ldots,I)P^H_{C_1}(K \mid N,\ldots,L)P^H_{C_0}(D \mid N,\ldots,K)$$

FIG. 5. Evaluation of the α-helix segment model for a particular amino acid subsequence. Grayed areas are the N- and C-cap positions specified by distinct amino acid probability distributions. Internal positions are modeled as identically distributed but dependent. Throughout, amino acid distributions are conditioned on neighboring residues according to known helical side chain interactions as described in Section 2.2.

TABLE 1. SEGMENT CAPPING POSITIONS

Cap position	*Kullback-Leibler divergence from internal position*		
	α-helix	*β-strand*	*Loop/coil*
N1	**.146**	**.053**	**.050**
N2	**.196**	.034	**.032**
N3	**.115**	.025	.020
N4	**.08**	—	.014
C4	.019	—	.008
C3	.029	.020	.012
C2	.037	.015	.011
C1	**.059**	**.077**	**.056**

Kullback-Leibler divergence (cross-entropy) measured from the amino acid distribution at internal segment positions to N- and C-terminal positions. Positions shown in boldface are included in capping models as described in text. KL divergence between two probability distributions p and q is defined as $\sum_i p(i) \log\left(\frac{p(i)}{q(i)}\right)$.

end-capping residues in β-strands and loops is less obvious than in the case of α-helices. In choosing ℓ^E_N, ℓ^E_C, ℓ^L_N, ℓ^L_C we once again considered up to 4 positions for loops and 3 for β-strands (due to sparse data). Again, Table 1 shows the statistical deviance, and it is seen that β-strands and loops show little change in these positions. Accordingly, we set $\ell^E_N = 1$, $\ell^E_C = 1$, $\ell^L_N = 2$, $\ell^L_C = 1$. Figures 4A and 4B reveal some expected patterns in the associated amino acid distributions, such as Pro in position 2 as an initial[1] helix-terminating position in loops, a prevalence of Gly in internal loop positions, and various hydrophobic residues in strands.

Another difference between the models for α-helices, β-strands, and loops lies in the exact form of (4), reflecting the differing intra-segment correlations induced by the underlying backbone-side chain geometry being modeled. Reflecting the periodicity of β-strand side chains, conditioning is done on residues $i - 1$ and $i - 2$, and loops are modeled similarly.

[1] The occurrence of the peak at position 1 rather than position 2 is again an artifact of the helix boundaries defined by DSSP.

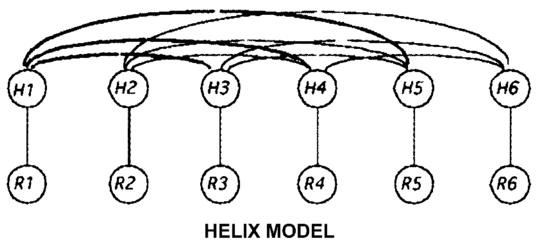


FIG. 6. A graphical model (Whittaker, 1990) representing the conditional independence structure for the amino acids in an example α-helix segment. $\mathrm{R_i}$ are the amino acids of the α-helix and $\mathrm{H_i}$ are their associated hydrophobicity classes as assigned by (Klingler and Brutlag, 1994). The model provides for dependence among the hydrophobicity classes at appropriate periodicity allowing the amino acid distributions to be modeled as *conditionally* independent, thus reducing the dimensionality of the model.

Finally, we note that no algorithmic model selection was done to select the best models of form (3) and (4), and so the models described above, while effective, are not necessarily optimal. Moreover, (3) and (4) are only one of many conceivable forms for the segment models, and many other possibilities exist. For example, many of the statistical models used for window-based prediction methods might be adapted for this purpose. Thus, we view the development of new *models* for structural segments to be a promising area of research. So long as the factorization given by (1) is maintained, our general framework holds and the computational methods described in the next section are applicable. Generalizations of (1) itself are discussed in Section 4.2.

2.3. *Computation and inference*

Assuming the probability model given by (1–4), we wish to infer the secondary structure assignment parameters (m, S, T) for a new protein sequence R. Thus we wish to find (m, S, T) such that $P(m, S, T \mid R)$ is maximized.[2] As mentioned in Section 2.1, our class of models is structurally similar to the class of semi-Markov source models described in Rabiner (1989). Thus, computation can be done exactly using a slight generalization of the standard forward-backward algorithm for hidden Markov models (HMMs), as described in Rabiner (1989), using the forward and backward variables defined as follows:

$$\begin{aligned}\alpha(j,t) = \sum_{v=l}^{j-1} \sum_{l \in SS} & \alpha(v,l) P(R_{[v+1:j]} \mid S_{prev} = v, S = j, T = t, \theta) \\ & \times P(S = j \mid T = t, S_{prev} = v, \theta) P(T = t \mid T_{prev} = l, \theta)\end{aligned} \tag{5}$$

$$\begin{aligned}\beta(j,t) = \sum_{v=j+1}^{n} \sum_{l \in SS} & \beta(v,l) P(R_{[j+1:v]} \mid S_{next} = v, S = j, T_{next} = l, \theta) \\ & \times P(S_{next} = v \mid S = j, T_{next} = 1, \theta) P(T_{next} = l \mid T = t, \theta)\end{aligned} \tag{6}$$

where $SS = \{H, E, L\}$, the set of possible secondary structural types and θ represents the model parameters. This yields an $O(n^3)$ algorithm, but in practice we limit the maximum size considered for any one segment to some length D. Thus, the first summation in (5) begins at $(j - D)$ and the first summation in (6) ends

[2] Throughout, the parameters of the probability model are assumed to be fixed, and we discuss only computation of predictive quantities of interest. Estimation of these probability parameters from the structural database described in Section 3 is straightforward using maximum likelihood or maximum a posteriori methods and amounts to counting observed frequencies for the desired quantities (the length of α-helices, or the occurrence of particular amino acids in the C-terminal capping position of an α-helix, for example). Because the database contains only sequences with known structures, no Baum-Welch type iteration is required during estimation. This contrasts with the use of HMM-like models in many other applications (such as multiple sequence alignment), where the underlying model is assumed unknown, or the data is not fully observable.

at $(j+D)$, yielding an algorithm which is linear $(O(nD^2))$ in the length of the input sequence for fixed D. All experiments in this paper use a value of $D = 30$, which is sufficiently large to account for nearly all observed structural segments as can be seen from Figure 2. We note that the model given by (3) allows further reduction of the D^2 term to yielding $O(nD)$; however, this does not hold in general and in practice the additional computational savings provided by this form are unnecessary.

We can therefore compute the *maximum a posteriori* values of (m, S, T):

$$Struct_{MAP} = \arg \max_{(m,S,T)} P(m, S, T \mid R, \theta)$$

using a procedure analogous to the Viterbi algorithm for HMMs (see Rabiner (1989)), simply by replacing the summations in (5) with maximization. We refer to these values of (m, S, T) as the *MAP segmentation*. A similar approach is taken by Burge and Karlin (1997) to find the optimal parse of a DNA sequence. We note, however, that many different segmentations may exist which, although not optimal, may have significant probability mass. In addition, the most commonly reported measure of accuracy for protein secondary structure prediction is the Q_3 value, the percentage correct on a *per residue* basis. Thus the MAP segmentation is not as desirable as the *marginal posterior mode* at each position:

$$Struct_{Mode} = \left\{ \arg \max_{T} P(T_{R_{[i]}} \mid R, \theta) \right\}_{i=1}^{n}$$

where $P(T_{R_{[i]}} \mid R, \theta)$ represents the marginal posterior distribution over structural types at position i. Fortunately, this is easily calculated from (5) and (6) above:

$$P(T_{R_i} = t \mid R, \theta) = \sum_{j=i-D+1}^{i-1} \sum_{k=i}^{j+D-1} \sum_{l \in SS} \alpha(j, l)\beta(k, t) P(R_{[j+1:k]} \mid S_{prev} = j, S = k, T = t, \theta) \\ \times P(S = k \mid S_{prev} = j, T = t, \theta) P(T = t \mid T_{prev} = l, \theta)/Z \tag{7}$$

where Z is the normalizing constant (or partition function) $P(R \mid \theta)$ which is available directly from the forward pass (5). The calculation in (7) yields the marginal posterior distribution at each position in the

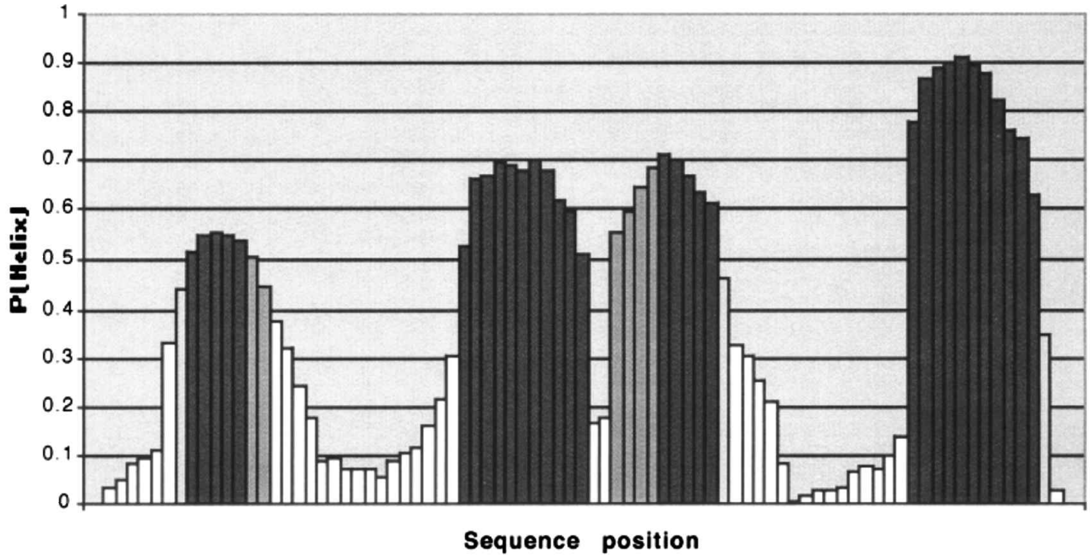


FIG. 7. Helix prediction probabilities for example sequence (1cc5). Positions in black (white) are correctly predicted by the BSPSS algorithm to be helix (coil); light gray positions are under-predicted (true structure helix, predicted structure coil), and dark gray are over-predicted (true structure coil, predicted structure helix). Prediction probabilities at each position correlate highly with prediction accuracy (see Figure 8).

sequence in $O(nD)$ time. We show in Section 3 that this marginal mode strategy significantly outperforms the MAP segmentation strategy by the Q_3 measure.

It is worth reiterating that (7) gives us the *exact* marginal posterior distribution over secondary structural types at each position, averaging over *all possible* segmentations, and hence provides an exact measure of the uncertainty of prediction at each position (subject to modeling assumptions). Figure 8 in Section 3 shows that this measure correlates very strongly with prediction accuracy and is still somewhat conservative. Figure 7 shows a typical sequence prediction, where we see that segment endpoints are the regions of highest uncertainty, as we would expect given the variability of assignments in such positions (Colloc'h *et al.*, 1993).

Lastly, we note that our approach can easily incorporate prior knowledge about regions or positions in the sequence if such is available. That is, the methods described in this section can be easily modified to calculate probabilities *conditional* on certain positions or segments taking on known conformations. Such might be the case, for example, if experimental evidence exists such as circular dichroism data or footprinting experiments, or if highly significant motif hits occur on the sequence and we wish to include them, for example, with helix-turn-helix DNA binding motifs. Again, such information is inherently difficult to include in most existing secondary structure prediction methods.

3. RESULTS

In order to evaluate the accuracy of our approach, we created a nonredundant set of 452 globular protein structures from the Protein Data Bank (Bernstein *et al.*, 1977) using OBSTRUCT (Heringa *et al.*, 1992). We created a maximal set of structures determined at better than 2.5 angstroms resolution with less than 25% sequence identity, removing those structures classified as membrane proteins within the SCOP hierarchy (Murzin *et al.*, 1995) and those sequences less than 50 amino acids in length. Table 2 reports the results of cross-validation experiments whereby each structure was predicted in turn, using parameters of (2), (3), (4) estimated from the remaining 451 structures. Quantities reported are the total percent correct (Q_3), percent of each structural type predicted correctly (sensitivity), percent of predictions for each type which were correct (positive predictive value), and Matthew's correlation (Matthews, 1975). Computation time on an SGI 195MHz Octane ranges from 0.2 seconds for the shortest sequence (50 residues) to 6.4 seconds for the longest (869 residues). The gold-standard secondary structure assignments were taken to be those provided by DSSP (Kabsch and Sander, 1983), with adjustments following (Frishman and Argos, 1996) to restrict the minimum β-strand length to 3, and the minimum α-helix length to 5. Our Bayesian segmentation algorithm (BSPSS) achieves a Q_3 accuracy of 68.8%, as high as most published single-sequence methods (Frishman and Argos, 1996) and only slightly below the best reported value of 71% (Salamov and Solovyev, 1997). Of the two predictors described in Section 2.3, the marginal mode at each position significantly outperforms the MAP segmentation on a per residue basis.

As described in Section 2.3, the BSPSS algorithm calculates the exact posterior distribution over structural types at each position. Figure 8 shows the Q_3 accuracy as a function of the probability assigned to the predicted structure at each position. As can be seen from the strong correlation, a clear advantage of our explicit probabilistic approach is the ability to accurately estimate the confidence in prediction at each position. At a threshold prediction probability of .6, we make predictions for 58% of positions and achieve an accuracy of 80.6%. At a threshold probability of .8 we achieve an accuracy of 91.4%, but predict only

Table 2. Accuracy of Secondary Structure Predictions

	Percent correct (%)				*Matthews' correlation*		
	Total	*Helix*	*Strand*	*Loop*	*Helix*	*Strand*	*Loop*
	(Q_3)	(Q_α)	(Q_β)	(Q_L)	(C_α)	(C_β)	(C_L)
MAP segmentation	64.2	67.3 (61.8)	23.3 (61.3)	79.1 (65.9)	.49	.30	.38
Marginal mode	68.8	64.0 (69.7)	46.0 (61.0)	81.0 (70.5)	.54	.43	.47

Results of experiments described in Section 3 evaluating the predictive accuracy for the two predictors defined in Section 2.3. Based on (1) the MAP segmentation and (2) the mode of the marginal posterior distribution at each position. Percentage correct is given as sensitivity (positive predictive value), and C_x are Matthews' correlation coefficients as defined in Matthews (1975).

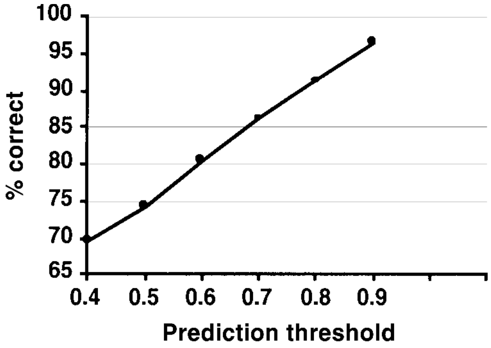


FIG. 8. Plot of predictive accuracy vs. probability assigned to prediction. The clear correlation indicates the value of prediction probabilities for interpreting structure predictions at various positions in the target sequence.

21% of positions with this level of confidence. It is worth noting that, according to Rost and Schneider (1997), these threshold percentages indicate that the BSPSS algorithm performs 6 times as well as other single sequence methods with reliability estimates, and multiple sequence alignment methods such as PhD (Rost and Sander, 1994) perform only 7/6 times better than BSPSS.

As described in Section 2, some model selection was done based on the dataset described (such as the determination of model cap lengths in Section 2.2). In principle, this means that the cross-validation results given for this dataset may potentially overestimate the accuracy expected for sequences outside this dataset. In order to account for this, we repeated the above experiments after model selection had been completed, using a new (larger) dataset which includes structures published in the interim. We used a dataset generated by the PDB_SELECT algorithm (Hobohm and Sander, 1994) with less than 25% sequence similarity. This set contained 685 protein structures, which was reduced to 660 by removal of structures classified as membrane proteins by SCOP (Murzin *et al.*, 1995) and those for which DSSP (Kabsch and Sander, 1983) produced no output. Results on this dataset were only slightly lower (68.4% for marginal mode predictions and 63.9% for MAP assignments), indicating that these estimates are largely unbiased. This difference in performance on the two datasets is comparable to that exhibited by other published algorithms when applied to multiple datasets (Frishman and Argos, 1997; Salamov and Solovyev, 1997).

4. DISCUSSION

4.1. Related work

In Section 2, we noted the similarity between our models and work in the speech recognition literature (Levinson, 1986; Rabiner, 1989; Russell and Moore, 1985) on semi-Markov source models. Here we briefly outline other related work.

Standard hidden Markov or Markov source models (Rabiner, 1989) have been used extensively in the literature on computational biology (Eddy, 1996; Krogh *et al.*, 1994). Recently, these have been related (Lawrence *et al.*, 1993; Liu *et al.*, 1999) in a unified framework to block-multinomial models for motif detection (Lawrence *et al.*, 1993; Liu *et al.*, 1995). HMMs have been applied to secondary structure prediction (Asai *et al.*, 1993; Stultz *et al.*, 1993) but have achieved limited accuracy. As described in detail in Section 2 above, we believe the intra-segment residue independence and geometric length distributions implied by HMMs to be inappropriate for modeling protein secondary structure, motivating our segment-based approach. Auger and Lawrence (1989) and Liu and Lawrence (1996) develop an algorithm for sequence segmentation which is analogous to the Viterbi algorithm for hidden semi-Markov models. Recent work on gene-finding in eukaryotic DNA sequences has used a similar approach (Burge and Karlin, 1997; Kulp *et al.*, 1996). Because of the linear nature of patterns in DNA, no attempt has been made to generalize these ideas to include segment interactions of the sort described in Section 4.2 below.

Although the vast majority of work on secondary structure prediction takes a window-based approach, the idea of locating segments of secondary structure goes back as far as the earliest work in this area (Chou and Fasman, 1974). More recently, Cohen *et al.* (1986) and Presnell *et al.* (1992) use deterministic pattern-matching methods to locate turns and helices, including regular expressions for helical caps. Solovyev and

Salamov (1994) use discriminant analysis of segment summary statistics (including residue pairs and the hydrophobic moment) to predict segments of secondary structure. Our approach is similar in spirit to many of these methods. Here, we place such ideas on firm theoretical ground within a Bayesian framework, providing explicit probabilistic models for protein sequence/structure relationships, and computational machinery for prediction and inference.

The α-helix models described in Section 2.2 contain as a special case statistical mechanical models developed in the theory of helix-coil transitions (Doig *et al.*, 1994; Poland and Scheraga, 1970; Stapley *et al.*, 1995). This connection and the implications of the more general framework provided here are explored in detail by Schmidler (2000).

4.2. Extensions and future directions

A fundamental assumption of (1) is the conditional independence of residues which occur in distinct segments. This assumption enables the exact calculation of posterior probabilities via (5–7). However, this assumption is clearly violated in the case of protein sequences. There exist numerous structural motifs which rely on sequentially distant segments interacting in three-dimensional space, including β sheets, coiled coils, disulfide bonds, and many others. The presence of correlated mutations in such motifs is well known; coiled coils may have stabilizing side chain interactions (Krylov *et al.*, 1994), and β-sheets may have charged-pair interactions and other correlations (Lifson and Sander, 1980; Smith and Regan, 1995), such as hydropathy correlations between strands induced by the environment of each face of the sheet. It has been frequently suggested that the inability to capture such nonlocal patterns in window-based classifiers may be responsible for the low accuracy typically achieved in β-strand prediction, and recent empirical results lend support to this view (Frishman and Argos, 1996). Only recently have attempts been made to incorporate such information into general prediction schemes with some success (Frishman and Argos, 1996). It is interesting that our BSPSS algorithm achieves accuracy slightly higher than the 68% reported by Frishman and Argos (1996) without consideration of such nonlocal interactions. Thus the inclusion of nonlocal information into our model may further improve accuracy.

Conceptually, it is straightforward to incorporate segment interaction terms into our model. For two segments j and k we need simply replace the two terms:

$$P(S_j \mid S_{j-1}, T_j)P(R_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, T_j) \quad \text{and} \quad P(S_k \mid S_{k-1}, T_k)P(R_{[S_{k-1}+1:S_k]} \mid S_{k-1}, S_k, T_k)$$

in the product of (1) and (2) above with a single term:

$$P(S_j, S_k \mid S_{j-1}, S_{k-1}, T_j, T_k)P(R_{[S_{j-1}+1:S_j]}, R_{[S_{k-1}+1:S_k]} \mid S_{j-1}, S_j, T_j, S_{k-1}, S_k, T_k) \tag{8}$$

for appropriate T_j, T_k. This enables us to incorporate arbitrary joint segment distributions, or *pairwise potentials*, into our model. The extension to three or more segments (e.g., for 4-helix bundles or amphipathic β-sheets) is obvious. However, the difficulty arises in calculation of the posterior distribution under such a model. As mentioned, the conditional independence exhibited by (1) is critical for recursive definition of the joint probability distribution over (R, m, S, T) and therefore for efficient calculation of posterior probabilities via dynamic programming. Although terms such as (8) can be introduced in a limited fashion via higher-order Markovian dependence in (1) and (2), the computational expense increases dramatically. Furthermore, it is not possible to allow interaction between arbitrary segments along the sequence in this fashion, without allowing Markovian dependence of order n. It is therefore infeasible to compute exactly the posterior distribution as was done in Section 2.3 above. We are currently developing efficient Monte Carlo sampling-based methods for approximate computation of the posterior distribution under such models (Schmidler, 2000; Schmidler *et al.*, 2000).

Significant evidence also exists that the inclusion of multiple sequence alignment information, when available, can improve single sequence prediction methods by as much as 5–7% (Di Francesco *et al.*, 1996; Rost and Sander, 1993a, 1994; Salamov and Solovyev, 1995). Indeed, the most accurate methods published to date (Frishman and Argos, 1997) utilize multiple sequence alignments to achieve improved performance. Extension of our algorithm to account for multiple aligned sequences is straightforward, and we are currently doing so (Schmidler, 2000).

5. CONCLUSION

We have presented a novel approach to the prediction of protein secondary structure from sequence. We provide probabilistic models for protein structural segments and an algorithm for prediction based on Bayesian inference. Evaluation of this approach on a database of 452 nonhomologous sequences with known structure achieves accuracies comparable to the best published single sequence methods, with the advantage of accurate estimates of prediction uncertainty. Extensions to more general models for segment interactions are discussed.

ACKNOWLEDGMENTS

We thank Chip Lawrence for helpful discussions. S.C.S. is supported by NLM training grant LM-07033. J.S.L. is partially supported by NSF grant DMS-9803649 and DOE grant DE-FG02-96ER62266. D.L.B. is supported by NLM grant LM-05716.

REFERENCES

Asai, K., Hayamizu, S., and Handa, K.I. 1993. Prediction of protein secondary structure by the hidden Markov model. *Comp. Applic. Biosci.* 9(2), 141–146.

Auger, I.E., and Lawrence, C.E. 1989. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* 51(1), 39–54.

Aurora, R., and Rose, G.D. 1998. Helix capping. *Protein Science* 7, 21–38.

Barton, G.J. 1995. Protein secondary structure prediction. *Curr. Opin. Struct. Biol.* 5, 372–376.

Benner, S.A. 1995. Predicting the conformation of proteins from sequences: progress and future progress. *J. Mol. Recogn.* 8, 9–28.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macro-molecular structures. *J. Mol. Biol.* 112, 535–542.

Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94.

Chou, P.Y., and Fasman, U.D. 1974. Prediction of protein conformation. *Biochemistry* 13, 211–215.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., and Fletterick, R.J. 1986. Turn prediction in proteins using a pattern-matching approach. *Biochemistry* 25, 266–275.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.-P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* 6(4), 377–382.

Di Francesco, V., Garnier, J., and Munson, P.J. 1996. Improving protein secondary structure prediction with aligned homologous sequences. *Protein Science* 5, 106–113.

Doig, A.J., and Baldwin, R.L. 1995. N- and C-capping preferences for all 20 amino acids in α-helical peptides. *Protein Science* 4, 1325–1336.

Doig, A.J., Chakrabartty, A., Klingler, T.M., and Baldwin, R.L. 1994. Determination of free energies of N-capping in α-helices by modification of the Lifson-Roig helix-coil theory to include N- and C-capping. *Biochemistry* 33, 3396–3403.

Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* 6, 361–365.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* 81, 140–144.

Fischer, D., and Eisenberg, D. 1996. Protein fold recognition using sequence-derived predictions. *Protein Science* 5, 947–955.

Friesner, R.A., and Gunn, J.R. 1996. Computer simulation of protein folding. *Ann. Rev. Biol. Biomol. Struct.* 25, 315–342.

Frishman, D., and Argos, P. 1996. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* 9(2), 133–142.

Frishman, D., and Argos, P. 1997. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* 27, 329–335.

Garnier, J., Gibrat, J.-F., and Robson, B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 266, 540–553.

Garnier, J., Osguthorpe, D.J., and 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.

Heringa, J., Sommerfeldt, H., Higgins, D., and Argos, P. 1992. OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *CABIOS* 8(6), 599–600.

Hobohm, U., and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* 3, 522–524.

Holley, H.L., and Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. USA* 86, 152–156.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33, 3038–3049.

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

Klingler, T.M., and Brutlag, D.L. 1994. Discovering structural correlations in α-helices. *Protein Science* 3, 1847–1857.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol*. 235, 1501–1531.

Krylov, D., Mikhailenko, I., and Vinson, C. 1994. A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. *EMBO Journal* 13(12), 2849–2861.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Fourth Intntl. Conf. Intell. Sys. Mol. Biol*.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.

Levinson, S.E. 1986. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language* 1, 29–45.

Lifson, S., and Sander, C. 1980. Specific recognition in the tertiary structure of β-sheets of proteins. *J. Mol. Biol*. 139, 627–639.

Liu, J.S., and Lawrence, C.E. 1996. Unified Gibbs method for biological sequence analysis. *Amer. Statist. Assoc., Statist. Comp. Section*.

Liu, J.S., Neuwald, A., and Lawrence, C.E. 1999. Markovian structures in biological sequence alignments. *J. Amer. Statist. Assoc*. 94, 1–15.

Liu, J.S., Neuwald, A.F., and Lawrence, C.E. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc*. 90, 1156–1170.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451.

Monge, A., Friesner, R.A., and Honig, B. 1994. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA* 91, 5027–5029.

Munson, P.J., Di Francesco, V., and Porrelli, R. 1994. Protein secondary structure prediction using periodic-quadratic-logistic models: Statistical and theoretical issues. *Twenty-seventh annual Hawaii international conference on system sciences.*

Murzin, A.Z., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. 247, 536–540.

Poland, D., and Scheraga, H.A. 1970. *Theory of helix-coil transitions in biopolymers*, Academic Press.

Presnell, S.R., Cohen, B.I., and Cohen, F.E. 1992. A segment-based approach to protein secondary structure prediction. *Biochemistry* 31, 983–993.

Presta, L.G., and Rose, G.D. 1988. Helix signal in proteins. *Science* 240, 1632–1641.

Qian, N., and Sejnowski, T.J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol*. 202, 865–884.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2), 257–286.

Richardson and Richardson. 1988. Amino acid preferences for specific locations at the ends of α-helices. *Science* 240, 1648–1652.

Riis, S.K., and Krogh, A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol*. 3(1), 163–183.

Rost, B., Fariselli, P., and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science* 5, 1704–1718.

Rost, B., and Sander, C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA* 90, 7558–7562.

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

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

Rost, B., and Schneider, R. 1997. Pedestrian guide to analysing sequence databases. In *Core techniques in biochemistry* Ashman, K., ed. Springer, Heidelberg.

Russell, M.J., and Moore, R.K. 1985. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. *ICASSP, Tampa, FL.*

Russell, R.B., Copley, R.R., and Barton, G.J. 1996. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol*. 259, 349–365.

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

Salamov, A.A., and Solovyev, V.V. 1997. Protein secondary structure prediction using local alignments. *J. Mol. Biol*. 268, 31–36.

Schmidler, S.C. 2000. Statistical models and Monte Carlo methods for protein structure prediction. Ph.D. thesis (in preparation), Stanford University.

Schmidler, S.C., Liu, J.S., and Brutlag, D.L. 2000. A Bayesian approach to predicting non-local interactions in protein sequences. *Submitted for publication*.

Smith, C.K., and Regan, L. 1995. Guidelines for protein design: the energetics of β-sheet side chain interactions. *Science* 270, 980–982.

Solovyev, V.V., and Salamov, A.A. 1994. Predicting α-helix and β-strand segments of globular proteins. *Comput. Appl. Biosci*. 10(6), 661–669.

Stapley, B.J., Rohl, C.A., and Doig, A.J. 1995. Addition of side chain interactions to modified Lifson-Roig helix-coil theory: Application to energetics of phenylalanine-methionine interactions. *Protein Science* 4, 2383–2391.

Stolorz, P., Lapedes, A., and Xia, Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol*. 225, 363–377.

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

Whittaker, J. 1990. *Graphical models in applied multivariate statistics*, Wiley.

Yi, T.-M., and Lander, E.S. 1993. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol*. 232, 1117–1129.

Zhang, X., Mesirov, J.P., and Waltz, D.L. 1992. Hybrid system for protein secondary structure prediction. *J. Mol. Biol*. 225, 1049–1063.

Address correspondence to:
Scott C. Schmidler
Section on Medical Informatics
Medical School Office Building, X215
Stanford University School of Medicine
Stanford, CA 94305

E-mail: Schmidler@SMI.stanford.edu