JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 2, 2006

Pp. 394–406

Protein Fold Recognition Using Segmentation Conditional Random Fields (SCRFs)

YAN LIU,[1] JAIME CARBONELL,[1] PETER WEIGELE,[2] and
VANATHI GOPALAKRISHNAN[3]

ABSTRACT

Protein fold recognition is an important step towards understanding protein three-dimensional structures and their functions. A conditional graphical model, i.e., segmentation conditional random fields (SCRFs), is proposed as an effective solution to this problem. In contrast to traditional graphical models, such as the hidden Markov model (HMM), SCRFs follow a discriminative approach. Therefore, it is flexible to include any features in the model, such as overlapping or long-range interaction features over the whole sequence. The model also employs a convex optimization function, which results in globally optimal solutions to the model parameters. On the other hand, the segmentation setting in SCRFs makes their graphical structures intuitively similar to the protein 3-D structures and more importantly provides a framework to model the long-range interactions between secondary structures directly. Our model is applied to predict the parallel β-helix fold, an important fold in bacterial pathogenesis and carbohydrate binding/cleavage. The cross-family validation shows that SCRFs not only can score all known β-helices higher than non-β-helices in the Protein Data Bank (PDB), but also accurately locates rungs in known beta-helix proteins. Our method outperforms BetaWrap, a state-of-the-art algorithm for predicting beta-helix folds, and HMMER, a general motif detection algorithm based on HMM, and has the additional advantage of general application to other protein folds. Applying our prediction model to the Uniprot Database, we identify previously unknown potential β-helices.

Key words: protein structure prediction, fold recognition, graphical models.

1. INTRODUCTION

IT IS WIDELY BELIEVED THAT PROTEIN STRUCTURES reveal important information about protein functions. One key step towards modeling a tertiary structure is to identify how building blocks of secondary structures arrange themselves in space, i.e., the supersecondary structures or protein folds. There has been significant work on predicting some well-defined types of structural motifs or functional units, such as $\alpha\alpha$- and $\beta\beta$-hairpins (Murzin *et al.*, 1995; Orengo *et al.*, 1997; Karplus *et al.*, 1998; Durbin *et al.*, 1998). The task of protein fold recognition is the following: given a protein sequence and a particular fold or

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.
[2]Biology Department, Massachusetts Institute of Technology, Cambridge, MA 02139.
[3]Center for Biomedical Informatics, University of Pittsburgh, PA 15260.

supersecondary structure, predict whether the protein adopts the structural fold and if so, locate the exact positions of each component in the sequence.

The traditional approach for protein fold prediction is to search the database for homologs of sequences with known structures using PSI-BLAST (Altschul *et al.*, 1997) or match against an HMM profile built from sequences with the same fold by HMMER (Durbin *et al.*, 1998) or SAM (Karplus *et al.*, 1998). These approaches work well for short motifs with strong sequence similarities. However, there exist many important motifs or folds without clear sequence similarity and involving the long-range interactions, such as the folds in the beta class (Menke *et al.*, 2004). These cases necessitate a more powerful model, which can capture the structural characteristics of the protein fold without requiring sequence similarity. Interestingly, the protein fold recognition task parallels an emerging trend in the machine learning community, i.e., the prediction problem for *structured* data, which predicts the labels of each node in a graph given an observation with particular structures, for example webpage classification using the hyperlink graph or object recognition using grids of image pixels. The *conditional* graphical models prove to be one of the most effective tools for this kind of problem (Kumar and Hebert, 2003; Pinto *et al.*, 2003).

In fact, several graphical models have been applied to protein structure prediction. One of the earliest approaches to this problem has been applying simple hidden Markov models (HMMs) to protein secondary structure prediction and protein motif detection (Karplus *et al.*, 1998; Durbin *et al.*, 1998; Bystroff *et al.*, 2000); Delcher *et al.* (1993) introduced probabilistic causal networks for protein secondary structure modeling. Recently, Liu *et al.* (2004) applied conditional random fields (CRFs), a discriminative graphical model based on undirected graphs, for protein secondary structure prediction; Chu *et al.* (2004) extended the segmental semi-Markov model (SSMM) under the Baysian framework for predicting protein secondary structures.

The bottleneck for protein fold prediction is the identification of long-range interactions, which refers to the hydrogen bending between amino acids far apart within the linear polypeptide sequence but close in space. For example, they could be either two β-strands with hydrogen bonds in a parallel β-sheet or helix pairs in coupled helical motifs. Generative models, such as HMM or SSMM, assume a particular data-generating process, which makes it difficult to consider overlapping features and long-range interactions. Discriminative graphical models, such as CRFs, takes on a single residue or residues of fixed length as an observation variable. Thus, they fail to capture the features over a whole secondary structure element or the interactions between adjacent elements in 3-D, which may be distant in the primary sequence. To solve the problem, we propose segmentation conditional random fields (SCRFs), which retain all the advantages of original CRFs and at the same time can handle observations of variable length.

2. CONDITIONAL RANDOM FIELDS (CRFS)

Simple chain-structured graphical models, such as hidden Markov models (HMMs), have been applied to various problems. As a "generative" model, HMMs assume that the data are generated by a particular model and compute the joint distribution of the observation sequence $\mathbf{x}$ and state sequence $\mathbf{y}$, i.e., $P(\mathbf{x}, \mathbf{y})$. However, generative models might perform poorly with inappropriate assumptions. In contrast, discriminative models, such as neural networks and support vector machines (SVMs), estimate the decision boundary directly without computing the underlying data distribution and thus often achieve better performance.

Recently, several discriminative graphical models have been proposed by the machine learning community, such as maximum entropy Markov models (MEMMs) (McCallum *et al.*, 2000) and conditional random fields (CRFs) (Lafferty *et al.*, 2001). Among these models, CRFs proposed by Lafferty *et al.*, are very successful in many applications, such as information extraction, image processing, and so on (Pinto *et al.*, 2003; Kumar and Hebert, 2003).

CRFs are "undirected" graphical models (also known as *random fields*, as opposed to directed graphical models such as HMMs) to compute the conditional likelihood $P(\mathbf{y}|\mathbf{x})$ directly. By the Hammersley–Clifford theorem (1971), the conditional probability $P(\mathbf{y}|\mathbf{x})$ is proportional to the product of the potential functions over all the cliques in the graph; that is,

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_0} \prod_{c \in C(\mathbf{y},\mathbf{x})} \Phi_c(\mathbf{y}_c, \mathbf{x}_c),$$

where $\Phi_c(\mathbf{y}_c, \mathbf{x}_c)$ is the potential function over the clique c, and Z_0 is the normalization factor over all possible assignments of $\mathbf{y}$ (see Jordan [1998] for more detail). For a chain structure, CRFs define the conditional probability as

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, i, y_{i-1}, y_i)\right), \tag{1}$$

where f_k is an arbitrary feature function over $\mathbf{x}$, N is the number of observations, and K is the number of features. The model parameters λ_k are learned via maximizing the conditional likelihood of the training data.

CRFs define the clique potential as an exponential function, which results in a series of nice properties. First, the optimization function is convex so that finding the global optimum is guaranteed (Lafferty *et al.*, 2001). Second, the feature definition can be arbitrary, including overlapping features and long-range interactions. Finally, CRFs still have efficient inference algorithms, such as forward–backward or Viterbi algorithm, as long as the graph structures are chains or trees.

Similarly to HMMs, we can define the forward-backward probability for CRFs. For a chain structure, the "forward value" $\alpha_i(y)$ is defined as the probability of being in state y at time i *given* the observation up to i. The recursive step for computing $\alpha_i(y)$ is

$$\alpha_{i+1}(y) = \sum_{y'} \alpha_i(y') \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, i+1, y', y)\right).$$

Similarly, $\beta_i(y)$ is the probability of starting from state y at time i *given* the observation sequence after time i. The recursive step is

$$\beta_i(y') = \sum_{y} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, i+1, y', y)\right)\beta_{i+1}(y).$$

The forward–backward algorithm and Viterbi algorithm can be derived accordingly (Sha and Pereira, 2003).

3. SEGMENTATION CONDITIONAL RANDOM FIELDS (SCRF'S)

Protein folds are frequent arrangement pattern of several secondary structure elements: some elements are quite conserved in sequences or prefer a specific length, while others might form hydrogen bonds with each other, such as two β-strands in a parallel β-sheet. To model the protein fold better, it would be natural to think of each secondary structure element as one observation, corresponding to one node in the graph, and the edges between elements as indicating their interactions in 3-D. Then, given a protein sequence, we can search for the best segmentation defined by the graph and determine whether the protein adopts the fold or not.

3.1. Protein structural graph

Before covering the algorithm in detail, we first introduce a special kind of graph, which we call the protein structural graph. Given a protein fold, a structural graph is defined as $G = \langle V, E_1, E_2 \rangle$, where $V = U \bigcup \{I\}$, U is the set of nodes corresponding to the secondary structure elements within the fold, and I is the node to represent the elements outside the fold. Set E_1 is the set of edges between neighboring elements in primary sequences, and E_2 is the set of edges indicating the potential long-range interactions between elements in tertiary structures. Figure 1 shows an example of the structural graph for β-α-β motif. Notice that there is a clear distinction between edges in E_1 and those in E_2 in terms of probabilistic semantics: similarly to HMMs, the E_1 edges indicate state transitions between adjacent nodes. On the other hand, the E_2 edges are used to model the long-range interactions, which is unique to the structural graph.

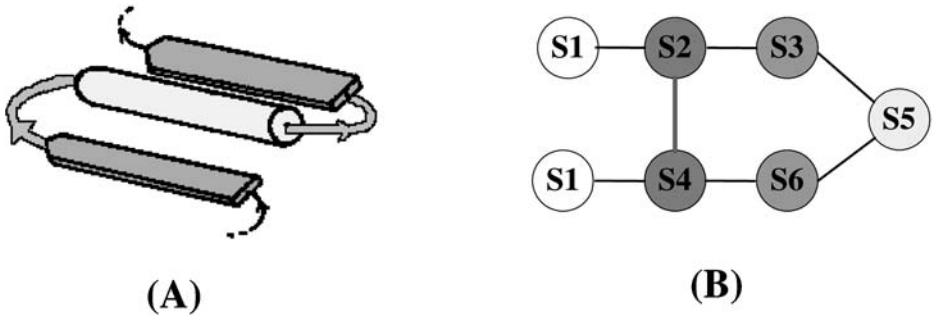


FIG. 1. Graph structure of β-α-β motif. (**A**) 3-D structure. (**B**) Protein structure graph. State set $= \{S_1, \ldots, S_6\}$, where S_1 represents non-β-α-β (I-node), S_2, S_4 represents β-strand, S_3, S_6 represents coil, S_5 represents α-helix. $E_1 = \{\text{black edges}\}$ and $E_2 = \{\text{grey bold edges}\}$.

In practice, one protein fold might correspond to several reasonable structural graphs given different semantics for each node. There is always a tradeoff between the graph complexity, fidelity of the model, and the real computational costs. Therefore, a good graph is the most expressive one that captures the properties of the protein folds while retaining as much simplicity as possible. There are several ways to simplify the graph; for example, we can combine multiple nodes with similar properties into one or remove those E_2 edges that are likely to be less predictive or less interesting to us. We give a concrete example of β-helix fold in Section 4.

3.2. Segmentation conditional random fields

Given a structural graph G and a protein sequence $\mathbf{x} = x_1 x_2 \ldots x_N$, we can have a possible segmentation of the sequence, i.e., $\mathbf{s} = (s_1, s_2, \ldots, s_M)$, where M is the number of segments, $s_i = \langle p_i, q_i, y_i \rangle$ with a starting position p_i, an end position q_i, and the state label of the segment y_i. The conditional probability of a segmentation $\mathbf{s}$ given the observation $\mathbf{x}$ can be computed as follows:

$$P(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_0} \prod_{c \in \mathcal{C}^G} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}_c, \mathbf{s}_c) \right),$$

where Z_0 is the normalization factor based on all possible configurations.

Since a protein fold is a regular arrangement of its secondary structure elements, the general topology is often known a priori and we can easily define a structural graph with deterministic transitions between adjacent nodes. Therefore, it is not necessary to consider the effect of E_1 edges in the model explicitly, resulting in a graph $G' = \langle V, E_2 \rangle$. If each subgraph of G' is a chain or a tree (an isolated node can also be seen as a chain), then we have

$$P(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^{M} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, s_i, s_{\pi_i}) \right), \tag{2}$$

where s_{π_i} is the predecessor (neighbor of smaller position index) of s_i in graph G'.

We estimate the parameters λ_k by maximizing the conditional log likelihood of the training data:

$$L_\Lambda = \sum_{i=1}^{M} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, s_i, s_{\pi_i}) - \log Z_0 + \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2},$$

where the last term is a Gaussian prior over the parameters as a smoothing term to deal with any sparsity problem in the training data. To perform the optimization, we need to seek the zero of the first derivative, i.e.,

$$\frac{\partial L_\Lambda}{\partial \lambda_k} = \sum_{i=1}^{M} (f_k(\mathbf{x}, s_i, s_{\pi_i}) - \mathrm{E}_{P(\mathbf{S}|x)}[f_k(\mathbf{x}, S_i, S_{\pi_i})]) + \frac{\lambda_k}{\sigma^2}, \tag{3}$$

where $\mathrm{E}_{P(\mathbf{S}|\mathbf{x})}[f_k(x, S_i, S_{\pi_i})]$ is the expectation of feature $f_k(\mathbf{x}, S_i, S_{\pi_i})$ over all possible segmentations of x. The convexity property guarantees that the root of Equation (3) corresponds to the optimal solution. However, since there is no closed-form solution, it is not straightforward to find the optimal solution. Recent work on iterative searching algorithms for CRFs suggests that L-BFGS converges much faster than other commonly used methods, such as iterative scaling or conjugate gradient (Sha and Pereira, 2003), which is also confirmed in our experiments for SCRFs.

Similarly to CRFs, we still have an efficient inference algorithm as long as each subgraph of G' is a chain and the nodes connected by E_2 edges have *fixed length* of residues. We redefine the forward probability $\alpha_{\langle l, y_l\rangle}(r, y_r)$ as the conditional probability that a segment of state y_r ends at position r given the observation $x_{l+1} \ldots x_r$ and a segment of state y_l ends at position l. The recursive step can be written as

$$\alpha_{\langle l, y_l\rangle}(r, y_r) = \sum_{p,\ p',\ q'} \alpha_{\langle l, y_l\rangle}(q', y')\alpha_{\langle q', y'\rangle}(p-1, \overleftarrow{y_r}) \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, s, s_\pi)\right),$$

where s_π is the predecessor of s in graph G'; i.e., $s = \langle p, r, y_r\rangle$ and $s_\pi = \langle p', q', y'\rangle$, "$\rightarrow$" is the operator to get the next state and "$\leftarrow$" the previous state (the value is known since the state transition is deterministic). The range over the summation is $\sum_{p=r-\ell_2+1}^{r-\ell_1+1} \sum_{q'=l+\ell_1'-1}^{p-1} \sum_{p'=l}^{q'-\ell_1'+1}$, where $\ell_1 = \max \text{length}(y_r)$, $\ell_2 = \min \text{length}(y_r)$. Then, the normalizer $Z_0 = \alpha_{\langle 0, y_{\mathtt{start}}\rangle}(N, y_{\mathtt{end}})$. Figure 2 shows a toy example of how to calculate the forward probability in detail.

Similarly, we can define the backward probability $\beta_{\langle r, y_r\rangle}(l, y_l)$ as the probability of $x_{l+1} \ldots x_r$ given a segment of state y_l ends at l and a segment of state y_r ends at r. Then we have

$$\beta_{\langle r, y_r\rangle}(l, y_l) = \sum_{q',\ p,\ q} \beta_{\langle r, y_r\rangle}(p-1, \overleftarrow{y}) \beta_{\langle p-1,\ \overleftarrow{y}\ \rangle}(q', \overrightarrow{y_l}) \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, s, s_\pi)\right),$$

where $s = \langle p, q, y\rangle$, $s_\pi = \langle l+1, q', \overrightarrow{y_l}\rangle$. Given the backward and forward algorithm, we can compute the expectation of each feature f_k in (3) accordingly.

For a test sequence, we search for the segmentation that maximizes the conditional likelihood $P(\mathbf{s}|\mathbf{x})$. Similarly to CRFs, we define

$$\delta_{\langle l, y_l\rangle}(r, y_r) = \max_{p,\ p',\ q'} \delta_{\langle l, y_l\rangle}(q', y')\delta_{\langle q', y'\rangle}(p-1, \overleftarrow{y_r}) \exp\left(\sum_k \lambda_k f_k(\mathbf{x}, s, s_\pi)\right).$$

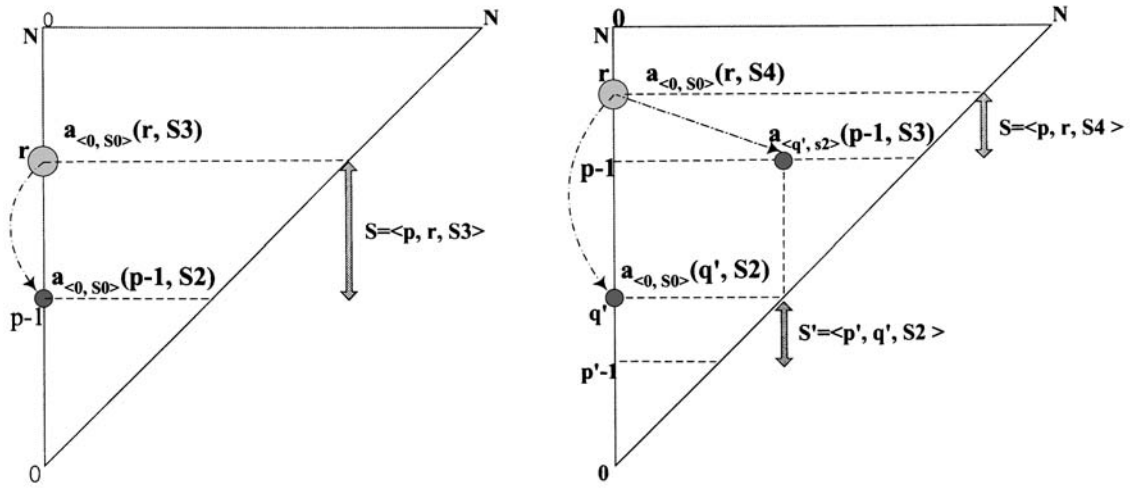


FIG. 2. An example of forward algorithm for the graph defined in Fig. 1B. The x/y-axis: index of starting/end residue position; light grey circle: target value; dark grey circle: intermediate value. (**Left**) calculation for $\alpha_{\langle 0, S_0\rangle}(r, S3)$ for segment S_3 with no direct forward neighbor; (**right**) calculation for $\alpha_{\langle 0, S_0\rangle}(r, S4)$ for segment S_4 with direct forward neighbor S_2.

The best segmentation can be traced back from $\delta_{\langle 0, y_{\mathtt{start}}\rangle}(N, y_{\mathtt{end}})$, where N is the number of residues in the sequence.

In general, the computational cost of SCRFs for the forward–backward probability and Viterbi algorithm will be polynomial to the length of the sequence N. However, in most real applications of protein fold prediction, the number of possible residues in each node is much smaller than N or fixed, which might reduce the final complexity to be approximately $O(N)$.

3.3. SCRFs for general graphs

For a general protein structural graph $G = \langle V, E_1, E_2\rangle$, the conditional probability of a sequence $\mathbf{x}$ given a segmentation $\mathbf{s}$ is defined as

$$P(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_0} \prod_{c \in \mathcal{C}^G} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}_c, s_c)\right).$$

If there are no long-range interactions involved within the fold, i.e., $E_2 = \emptyset$, SCRFs degrade into semi-Markov CRF models, which are linear CRFs that allow observations of variable length (Sarawagi and Cohen, 2004); on the other hand, if the state transitions between neighboring segments are deterministic, we do not need to consider the effect of E_1 edges in the graph, i.e., $E_1 = \emptyset$, which is the case that we described in detail in previous section. If neither E_1 nor E_2 is empty, the problem can be generalized as a structural learning problem for graphical models, and more thorough investigation is warranted.

For many protein folds or supersecondary structures, it is easy to construct a protein structural graph without loops and then exact inference algorithms can be applied; otherwise, approximation methods, for example, mean field approximation or loopy belief propagation, have to be applied (see Jordan [1998] for a good review).

4. APPLICATION TO RIGHT-HANDED PARALLEL β-HELIX PREDICTION

The right-handed parallel β-helix fold is an elongated helix-like structure with a series of progressive stranded coilings (called *rungs*), each of which is composed of three parallel β-strands to form a triangular prism shape (Yoder *et al.*, 1993). The typical 3-D structure of a β-helix is shown in Fig. 3A–B (Cowen *et al.*, 2002). As we can see, each basic structural unit, i.e., a rung, has three β-strands of various lengths, ranging from three to five residues. The strands are connected to each other by loops with distinctive features. One of the loops is a unique two-residue turn which forms an angle of approximately 120$^\diamond$ between two parallel β-strands (called the *T-2 turn*). The other two loops vary in size and conformation, which might contain coils, helixes, or even β-sheets. There currently exist 14 protein sequences with a three-stranded right-hand β-helix whose crystallized structures have been deposited in the Protein Data Bank (PDB) (See Table 1).

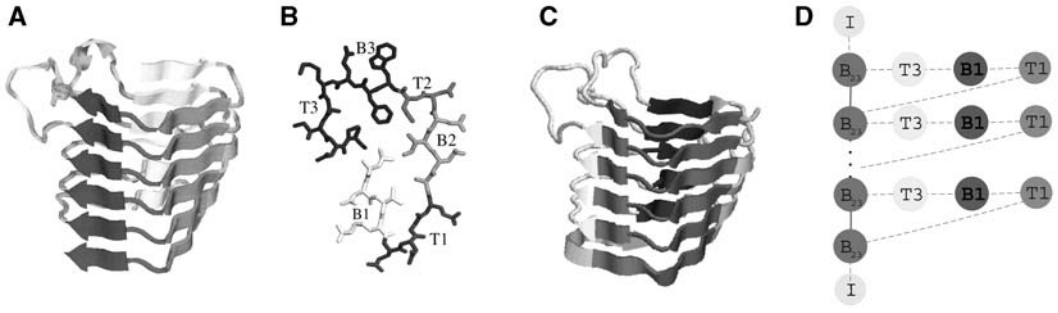


FIG. 3. Three-D structures and side-chain patterns of β-helices; (**A**) side view, (**B**) top view of one rung, (**C**) segmentation of 3-D structures, (**D**) protein structural graph. E1 = {dash lines}, and E2 = {solid lines} (Figs. (A) and (B) are adapted from Cowen *et al.* [2002]).

TABLE 1. SCORES AND RANK FOR THE KNOWN RIGHT-HANDED β-HELICES BY HMMER, BETAWRAP, AND SCRFs[a]

SCOP family	PDB ID	Struct-based bit score[b]	HMMs rank	Seq-based bit score[b]	HMMs rank	BetaWrap[a] Score	BetaWrap[a] Rank	SCRFs ρ-Score	SCRFs Rank
P.69 pertactin	1dab	−73.6	3	−163.4	75	−17.84	1	10.17	1
Chondroitinase B	1dbg	−64.6	5	− 171.0	55	−19.55	1	13.15	1
Glutamate synthase	1ea0	−85.7	65	−109.1	72	−24.87	N/A	6.21	1
Pectin methylesterase	1qjv	−72.8	11	−123.3	146	−20.74	1	6.12	1
P22 tailspike	1tyu	−78.8	30	−154.7	15	−20.46	1	6.71	1
Iota-carrageenase	1ktw	−81.9	17	− 173.3	121	−23.4	N/A	8.07	1
Pectate lyase	1air	−37.1	2	−133.6	35	−16.02	1	16.64	1
	1bn8	180.3	1	−133.7	37	−18.42	3	13.28	2
	1ee6	−170.8	852	−219.4	880	−16.44	2	10.84	3
Pectin lyase	1idj	−78.1	14	−178.1	257	−17.99	2	15.01	2
	1qcx	−83.5	28	−181.2	263	−17.09	1	16.43	1
Galacturonase	1bhe	−91.5	18	−183.4	108	−18.80	1	20.11	3
	1czf	−98.4	43	−188.1	130	−19.32	2	40.37	1
	1rmg	−78.3	3	−212.2	270	−20.12	3	23.93	2

[a]The scores and rank from BetaWrap are taken from Bradley *et al.* (2001a) except 1ktw and 1ea0.
[b]The bit scores in HMMER are not directly comparable.

The β-helix structures are significant in that they include pectate lyases, which are secreted by bacterial pathogens during the infection of plants; the phage P22 tail-spike adhesion, which binds the O-antigen of *Salmonella typhimurium*; and the P.69 pertactin toxin from Bordetella pertussis, the cause of whooping cough. Therefore, it would be very interesting if we can accurately predict other unknown β-helix structure proteins, which may also have significant functions.

Traditional methods for protein family classification, such as threading, PSI-BLAST, and HMMs, fail to solve the β-helix recognition problem across different families (Cowen *et al.*, 2002). Recently, a computational method called BetaWrap, has been proposed to predict the β-helix specifically (Bradley *et al.*, 2001b, 2001a; Cowen *et al.*, 2002). The algorithm "wraps" the unknown sequences in all plausible ways and check the scores to see whether any wrap makes sense. The cross-validation results in the protein data bank (PDB) seem quite promising. However, the BetaWrap algorithm hand-codes some biological heuristic rules, which makes it prone to over-fit the known β-helix proteins and also hard to apply for other fold prediction tasks. We are, however, indebted to the BetaWrap efforts for identifying meaningful features, as discussed in Section 4.3.

4.1. Protein structural graph for β-helix

From previous literature on β-helix structures, there are two facts important for accurate prediction: 1) the β-strands of each rung have patterns of pleating and hydrogen bonding that are well conserved across the superfamily and 2) the interaction of the strand side-chains in the buried core are critical determinants of the fold (Yoder and Jurnak, 1995; Kreisberg *et al.*, 2000). Therefore we define the protein structural graph of β-helix as in Fig. 3D.

There are five states in the graph altogether, i.e., s-B23, s-T3, s-B1, s-T1, and s-I. The state s-B23 is a union of B2, T2, and B3 because these three segments are all highly conserved in pleating patterns and a combination of conserved evidence is generally much easier to detect. We fix the length of S-B23 and S-B1 as 8 and 3 respectively for two reasons: first, these are the numbers of residues shared by all known β-helices; second, it helps limit the search space and reduce the computational costs. The states s-T3 and s-T1 are used to connect s-B23 and s-B1. It is known that the β-helix structures will break if the insertion is too long. Therefore, we set the length of s-T3 and s-T1 in a range from 1 to 80. State s-I is the non-β-helix state, which refers to all those regions outside the β-helix structures in the protein. The edge denoted by the solid line between s-B23 is used to model the long-range interaction between adjacent β-strand pairs. For a protein without any β-helix structures, we define the protein structural graph as a single node of state s-I.

4.2. SCRFs for β-helix fold prediction

In Section 3.2, we made two assumptions in the SCRFs model: a) the state transition is deterministic and b) each subgraph of $G' = \langle V, E_2 \rangle$ is a chain or a tree. For the β-helix, we cannot directly define a structural graph with deterministic state transitions, since the number of rungs in a protein is unknown beforehand. In Fig. 3, it seems that the previous state of s-B23 can be either s-I or s-T1. However, notice that s-I can appear only at the beginning or the end of a sequence; therefore, s-I can be the previous state of s-B23 iff the previous segment starts at the first residue in the sequence. Similarly, s-I can be the next state of s-B23 iff the next segment ends at the last residue. Therefore, *the state transition is deterministic given the constraint we have for s-I.* As for assumption b), it is straightforward to see that graph G' consists of a chain and a set of isolated nodes. Therefore, the algorithm discussed in Section 3.2 can be applied directly.

To determine whether a protein sequence has the β-helix fold, we define the score ρ as the log ratio of the probability of the best segmentation to the probability of the whole sequence as one segment in a null state s-I, i.e., $\rho = \log \frac{\max_s P(\mathbf{s}|\mathbf{x})}{P(\langle 1, N, s-I \rangle|\mathbf{x})}$. The higher the score ρ, the more likely that the sequence has a β-helix fold. We did not consider the long-range interactions between B1 strands explicitly since the effect is relatively weak given only three residues in s-B1 segments. However, we use the B1 interactions as a filter in the Viterbi algorithm: specifically, $\delta_t(y)$ will be the highest value whose corresponding segmentation also has the alignment scores for B1 higher than a threshold set using cross-validation.

4.3. Feature extraction

SCRFs provide an expressive framework to handle long-range interactions for protein fold prediction. However, the choice of feature function f_k plays a key role in accurate predictions. We define two types of features for β-helix prediction, i.e., *node features* and *internode features*.

Node features cover the properties of an individual segment, including:

a. **Regular expression template:** Based on the side-chain alternating patterns in the B23 region, BetaWrap generates a regular expression template to detect a union of B2-T2-B3 strands, i.e., ΦXΦXXΨXΦX, where Φ matches any of the hydrophobic residues {A, F, I, L, M, V, W, Y}, Ψ matches any amino acids except ionisable residues {D, E, R, K}, and X matches any amino acid (Bradley *et al.*, 2001b). Following this pattern, we define the feature function $f_{RST}(\mathbf{x}, s_i)$ equal to 1 if the segment s_i matches the template, and 0 otherwise.

b. **Probabilistic HMM profiles:** The regular expression template above is straightforward and easy to implement. However, sometimes it is hard to make a clear distinction between a true motif and a false alarm. Therefore, we built a probabilistic motif profile using HMMER (Durbin *et al.*, 1998) for the s-B23 and s-B1 segments, respectively. We define the feature functions $f_{HMM1}(\mathbf{x}, s_i)$ and $f_{HMM2}(\mathbf{x}, s_i)$ as the alignment scores of s_i against the s-B23 and s-B1 profiles.

c. **Secondary structure prediction scores:** Secondary structures reveal significant information on how a protein folds in three dimensions. The state-of-art prediction method can achieve an average accuracy of 76–78% on soluble proteins. We can get fairly good prediction on alpha-helix and coils, which can help us locate the s-T1 and s-T3 segments. Therefore, we define the feature functions $f_{ssH}(\mathbf{x}, s_i)$, $f_{ssE}(\mathbf{x}, s_i)$, and $f_{ssC}(\mathbf{x}, s_i)$ as the average of the predicted scores by PSIPRED over all residues in segment s_i, for helix, sheet, and coil, respectively (Jones, 1999).

d. **Segment length:** It is interesting to notice that the β-helix structure has strong preferences for insertions within certain length ranges (see Fig. 4). To consider this preference in the model, we did parametric density estimation. We explore several common distribution functions, including Poisson distributions, negative-binomial distributions, and asymmetric exponential distributions, which consist of two reverse-polarity exponential functions meeting at one point. We use the latter one since it provides a better estimator than the other two (see Fig. 4). Then we define the feature functions $f_{L1}(\mathbf{x}, s_i)$ and $f_{L3}(\mathbf{x}, s_i)$ as the estimated density of length (s_i) under the distribution of length (s-T1) and length (s-T3) respectively.

Internode features capture long-range interactions between adjacent β-strand pairs, including:

a. **Side chain alignment scores:** BetaWrap calculates the alignment scores of residue pairs depending on whether the side chains are buried or exposed. In this method, the conditional probability that a residue

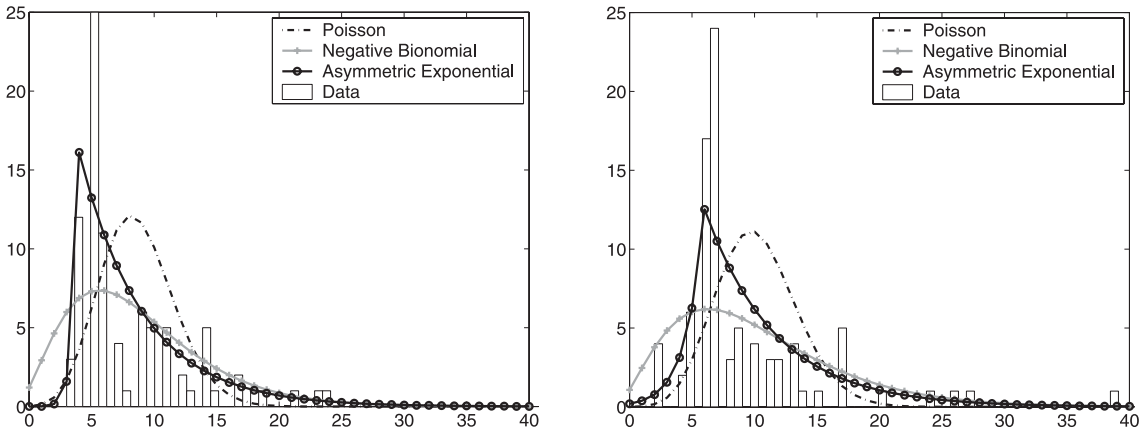


FIG. 4. Histograms for the length of s-T1 (**left**) and s-T3 (**right**).

of type X will align with residue Y, given their orientation relative to the core (buried or exposed), is estimated from a β-structure database developed from the whole PDB (Bradley *et al.*, 2001b). Following a similar idea, we define the feature function $f_{SAS}(\mathbf{x}, s_i, s_{\pi_i})$ as the weighted sum of the side chain alignment scores for s_i given s_{π_i} if both are s-B23 segments, where a weight of 1 is given to inward pairs and 0.5 to the outward pairs.

b. **Parallel β-sheet alignment scores:** In addition to the side chain position, another aspect to study is the different preferences of each amino acid to form parallel and anti-parallel β-sheets. Steward and Thornton (2002) derived the "pairwise information values" (V) for a residue of type X given the residue Y on the pairing parallel (or anti-parallel) strand and the offsets of Y from the paired residue Y′ of X. The alignment score for two segments $x = X_1 \dots X_m$ and $y = Y_1 \dots Y_m$ is defined as

$$score(x, y) = \sum_i \sum_j (V(X_i|Y_j, i-j) + V(Y_i|X_j, i-j)).$$

Unlike the side chain alignment scores, this score also takes into account the effect of neighboring residues on the paired strand. We define the feature function $f_{PAS}(\mathbf{x}, s_i, s_{\pi_i}) = score(s, s_{\pi_i})$ if the states of s_i and s_{π_i} are both s-B23 and 0 otherwise.

c. **Distance between adjacent s-B23 segments:** There are also different preferences for the distance between adjacent s-B23 segments. It is difficult to estimate this distribution since the range is too large. Therefore, we simply define the feature function as the normalized length; i.e., $f_{DIS}(x, s_i, s_{\pi_i}) = \frac{dis(S,S')-\mu}{\sigma}$, where μ is the mean and σ^2 is the variance.

Notice that most features defined above are quite general, not limited to predicting β-helices. For example, an important aspect to discriminate a specific protein fold from others is to build HMM profiles or identify regular expression templates for conserved regions if they exist; the secondary structure assignments are essential in locating the elements within a protein fold; if some segments have strong preferences for a certain length range, then the length is also informative. For internode features, the β-sheet alignment scores are useful for folds in the β-family while hydrophobicity is important for alpha or the alpha-beta class.

5. EXPERIMENTS

In our experiments, we followed the setup described by Bradley *et al.* (2001b). A PDB-minus dataset was constructed from the PDB protein sequences (July 2004 version) (Berman *et al.*, 2000) with less than 25% similarity to each other and no less than 40 residues in length. Then the β-helix proteins are removed from the dataset, resulting in 2,094 sequences in total. The proteins in the PDB-minus dataset will serve

as negative examples in the cross-family validation and later for discovery of new β-helix proteins. Since negative data dominate the training set, we subsample 15 negative sequences that are most similar to the positive examples in sequence identity so that SCRFs can learn a better decision boundary than randomly sampling.

5.1. Cross-family validation

A leave-family-out cross-validation was performed on the nine β-helix families of closely related proteins in the SCOP database (Murzin *et al.*, 1995). For each cross, proteins in one β-helix family are placed in the test set while the remainder are placed in the training set as positive examples. Similarly, the PDB-minus was also randomly partitioned into nine subsets, one of which is placed in the test set while the rest serve as the negative training examples. We compare our results with BetaWrap, the state-of-art algorithm for predicting β-helices, and HMMER, a general motif detection algorithm based on a simple graphical model, i.e., HMMs. The input to HMMER is a multiple sequence alignment. The best multiple alignments are typically generated using 3-D structural information, although this is not strictly a "by sequence alone" method. Therefore, we generated two kinds of alignments for comparison: one is multiple structural alignments using CE-MC (Guda *et al.*, 2004), the other is purely sequence-based alignments by CLUSTALW (Thompson *et al.*, 1994).

Table 1 shows the output scores by different methods and the relative rank of the β-helix proteins in the cross-family validation. From the results, we can see that the SCRFs model can successfully score all known β-helices higher than non-β-helices in the PDB. On the other hand, there are two proteins (i.e., 1ktw and 1ea0) in our validation sets that were crystallized recently and thus are not included in the BetaWrap system. We tested these two sequences on BetaWrap and got a score of -23.4 for 1ktw and -24.87 for 1ea0. These values are significantly lower than the scores of all other β-helices and even some of the non-β-helix proteins, which indicates that our method outperforms the BetaWrap algorithm. As expected, HMMER did worst even using the structural alignments.

Figure 5 plots the score histogram for known β-helix sequences against the PDB-minus dataset. Compared with the histograms in a similar experiment using BetaWrap (Bradley *et al.*, 2001b), our log ratio score ρ indicates a clear separation of β-helix proteins versus non-β-helix proteins. Only 18 out of 2,094 proteins have a score higher than zero. Among these 18 proteins, 13 proteins belong to the beta class and 5 proteins belong to the alpha-beta class in the CATH Database (Orengo *et al.*, 1997). In Table 2 we cluster the proteins into three different groups according to the segmentation results and include some examples of the predicted segmentation in each group.

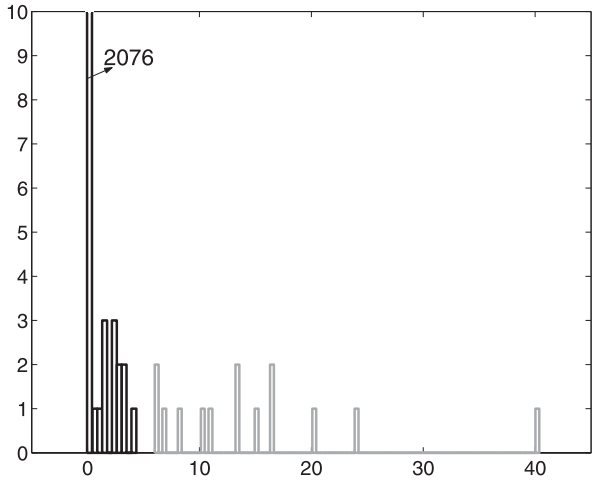
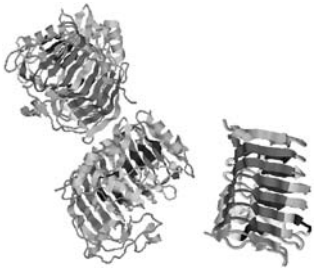
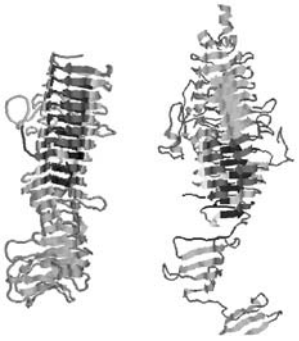


FIG. 5. Histograms of protein scores of known β-helix proteins against PDB-minus dataset. Dark grey bar: PDB-minus dataset; light grey bar: known β-helix proteins. Out of 2094 protein sequences in PDB-minus, 2076 have a log ratio score ρ of 0, which means that the best segmentation is a single segment in non-β-helix state.

TABLE 2. GROUPS OF SEGMENTATION RESULTS FOR THE KNOWN RIGHT-HANDED β-HELIX

Group	*Perfect match*	*Good match*	*OK match*
Missing rungs	*0*	*1–2*	*3 or more*
PDB ID	**1czf**	1air, 1bhe, 1bn8, 1dbg, **1ee6** (right), 1idj, **1ktw** (left), 1qcx, 1qjv, 1rmg	**1dab** (left), 1ea0, **1tyu** (right)
			

5.2. Discovery of potential β-helix proteins

New potential β-helix proteins were identified from the UniProt reference databases (UniRef) (a combination of Swiss-Prot Release 44.2 of 30 July 2004 and TrEMBL 27.2 of 30 July 2004) (Leinonen *et al.*, 2004). We choose the UniRef50 (50% identity) with 490,713 sequences as the exploration set. Ninety-three sequences were identified by the SCRFs model with scores above a cutoff of 5, all of which are identified as potential beta-helices. The sequences come from organisms in all domains of life. Of 44 eukaryotic sequences, 25 are from plants. The remaining eukaryotic sequences come from mammals, fungi, nematodes, and pathogens from the genus *Plasmodium*: four sequences were viral, including three from bacteriophages; nine sequences are archeal, seven of which are from methanogens of the genus *Methanosarcina*. Of the 93 high-scoring sequences, 48 are likely homologous (BLAST E-value < 0.001) with proteins currently known to contain parallel beta-helix domains. For the rest, most sequences are not homologous to any of the sequences in the PDB. The protein sequences with maximal log ratio scores is shown in Table 3, among which the polygalacturonases have already been shown to form parallel beta-helices, the CASH family of protein domains are also believed to be parallel beta-helices, and auto-transporter proteins may carry parallel beta-helices as a "passenger domain "; that is, the transporter domain exports the parallel beta-helix portion of the molecule to the outer cell surface. The full list can be accessed at *www.cs.cmu.edu/∼yanliu/SCRF.html*.

TABLE 3. EXAMPLES OF PROTEINS PREDICTED TO FORM β-HELIX IN UNIPROT

UniProt ID	*Description*	*Score*
Q8YK40	Auto-transporter/adhesin related	119.7
Q8PRX0	Cell surface glycoprotein/NosD domain	93.8
Q8WTU9	Ribonuclease	81.3
Q8DK34	Contains CASH domain, may bind carbohydrate	81.1
Q8RD81	Fibrocystin homolog, surface adhesion protein	55.1
O26812	S-layer protein/endopolygalacturonase domain	54.2
Q6LZ14	Hypothetical S-layer	43.8
P35338	Exopolygalacturonase precursor	42.2
Q6ZGA1	Putative polygalacturonase	41.6
Q9K1Z6	Putative outer membrane protein	40.8

Our method also identifies gp14 of *Shigella* bacteriophage Sf6 as having a parallel beta-helix domain, giving it a score of 15.63. This protein was not included in the UniRef50 dataset because it was incorrectly grouped with the P22 tail-spike protein (1tyu), which was used in the training dataset. These two proteins share homologous capsid binding domains at their N-termini, which are not parallel beta-helices, while their C-terminal domains do not have any sequence identity. An Sf6 gp14 crystal structure has recently been solved and shown to be a trimer of parallel β-helices (R. Seckler, personal communication). Therefore, SCRFs not only can identify homologous sequences to the known proteins, but also succeed in discovering proteins with significantly less sequence similarity.

6. DISCUSSION AND CONCLUSION

In Bradley *et al.* (2001b), BetaWrap was compared with other alternative methods, such as PSI-BLAST and Threader. We repeated their experiments and got similar results confirming that these methods fail to detect β-helix proteins accurately. Now it would be interesting to seek answers to the following questions: Why is β-helix prediction difficult for these commonly used methods? Why can the SCRFs model perform better?

It seems that the β-helix motif is hard to predict because there are long-range interactions in the β-helix fold. In addition, the structural properties unique to the β-helix are not reflected clearly in the sequences. For example, the conserved templates for the s-B23 segment also appear many times in non-β-helix proteins; the side chain alignment propensities in β-sheets are also shared by β-sheets in other structures, such as the β-sandwich. Therefore the commonly used methods based on sequence similarity, such as PSI-BLAST and HMMER, cannot perform well in this kind of task. However, a combination of both sequence and structure characteristics might help to identify more β-helices, which is one of the major reasons why BetaWrap and SCRFs work well. The difference between these two methods is BetaWrap searches the combination space by defining a series of heuristic rules while SCRFs search automatically by maximizing the conditional likelihood of the training data under a unified graphical model, which guarantees the solution to be global optimally. Therefore, the SCRFs model is more general and robust, even though it uses similar features as the BetaWrap method.

There are several directions to improve the SCRFs model, which are interesting both computationally and empirically. One is to extend the SCRFs model for predicting other protein folds, such as the leucine-rich repeats (LLR) or triple β-spirals. A second direction is to extend the current work for modeling protein dynamics, i.e., constructing more powerful graphical models to capture the dynamic constraints in the 3-D protein structures. The latter, however, will be a major undertaking.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0225656. We thank Jonathan King for his input and biological insights and anonymous reviewers for their comments.

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-blast: A new generation of protein database search programs. *Nucl. Acids Res.* 25(17), 3389–3402.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. 2000. The Protein Data Bank. *Nucl. Acids Res.* 28, 235–242.

Bradley, P., Cowen, L., Menke, M., King, J., and Berger, B. 2001a. Betawrap: Successful prediction of parallel beta-helices from primary sequence reveals an association with many microbial pathogens. *Proc. Natl. Acad. Sci.* 98, 14819–14824.

Bradley, P., Cowen, L., Menke, M., King, J., and Berger, B. 2001b. Predicting the beta-helix fold from protein sequence data. *Proc. 5th Ann. ACM RECOMB Conference*, 59–67.

Bystroff, C., Thorsson, V., and Baker, D. 2000. HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* 301(1), 173–190.

Chu, W., Ghahramani, Z., and Wild, D.L. 2004. A graphical model for protein secondary structure prediction. *Proc. Int. Conf. on Machine Learning (ICML '04)*, 161–168.

Cowen, L., Bradley, P., Menke, M., King, J., and Berger, B. 2002. Predicting the beta-helix fold from protein sequence data. *J. Comp. Biol.* 9, 261–276.

Delcher, A., Kasif, S., Goldberg, H., and Xsu, W. 1993. Protein secondary-structure modeling with probabilistic networks. *Int. Conf. on Intelligent Systems and Molecular Biology (ISMB '93)*, 109–117.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, London.

Guda, C., Lu, S., Sheeff, E., Bourne, P., and Shindyalov, I. 2004. CE-MC: A multiple protein structure alignment server. *Nucl. Acids Res.* In press.

Hammersley, J., and Clifford, P. 1971. *Markov Fields on Finite Graphs and Lattices*. Unpublished manuscript.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202.

Jordan, M.I. 1998. *Learning in Graphical Models*, MIT Press, Boston, MA.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14(10), 846–856.

Kreisberg, J., Betts, S., and King, J. 2000. Beta-helix core packing within the triple-stranded oligomerization domain of the p22 tailspike. *Protein Sci.* 9(12), 2338–2343.

Kumar, S., and Hebert, M. 2003. Discriminative random fields: A discriminative framework for contextual interaction in classification. *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 1150–1159.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th Int. Conf. on Machine Learning*, 282–289.

Leinonen, R., Diez, F., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. 2004. Uniprot archive. *Bioinformatics* 20(17), 3236–3237.

Liu, Y., Carbonell, J., Klein-Seetharaman, J., and Gopalakrishnan, V. 2004. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics* 20(17), 3099–3107.

McCallum, A., Freitag, D., and Pereira, F.C.N. 2000. Maximum entropy Markov models for information extraction and segmentation. *Proc. Int. Conf. on Machine Learning (ICML '00)*, 591–598.

Menke, M., Scanlon, E., King, J., Berger, B., and Cowen, L. 2004. Wrap-and-pack: A new paradigm for beta structural motif recognition with application to recognizing beta trefoils. *Proc. 8th ACM RECOMB Conference*, 298–307.

Murzin, A., Brenner, S., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247(4), 536–540.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* 5(8), 1093–1108.

Pinto, D., McCallum, A., Wei, X., and Croft, W.B. 2003. Table extraction using conditional random fields. *Proc. 26th ACM SIGIR Conference*, 235–242.

Sarawagi, S., and Cohen, W.W. 2004. Semi-Markov conditional random fields for information extraction. *Advances in Neural Information Processing Systems (NIPS 2004)*.

Sha, F., and Pereira, F. 2003. Shallow parsing with conditional random fields. *Proc. Human Language Technology (NAACL 2003)*.

Steward, R., and Thornton, J. 2002. Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins* 48(2), 178–191.

Thompson, J., Higgins, D., and Gibson, T. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22, 4673–4680.

Yoder, M., and Jurnak, F. 1995. Protein motifs. 3. The parallel beta helix and other coiled folds. *FASEB J.* 9(5), 335–342.

Yoder, M., Keen, N., and Jurnak, F. 1993. New domain motif: The structure of pectate lyase c, a secreted plant virulence factor. *Science* 260(5113), 1503–1507.

Address correspondence to:
Yan Liu
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

E-mail: yanliu@cs.cmu.edu