

Segmentation Conditional Random Fields (SCRFs): A New Approach for Protein Fold Recognition

Yan Liu¹, Jaime Carbonell¹, Peter Weigle², and Vanathi Gopalakrishnan³

¹ School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
{yanliu, jgc}@cs.cmu.edu

² Biology Department, Massachusetts Institute of Technology, Cambridge, MA, USA
pweigle@mit.edu

³ Center for Biomedical Informatics, University of Pittsburgh, PA, USA
vanathi@cbmi.pitt.edu

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 to solve the problem. In contrast to traditional graphical models such as hidden markov model (HMM), SCRFs follow a discriminative approach. It has the flexibility to include overlapping or long-range interaction features over the whole sequence, as well as global optimally solutions for the parameters. On the other hand, the segmentation setting in SCRFs makes its graphical structures intuitively similar to the protein 3-D structures and more importantly, provides a framework to model the long-range interactions directly.

Our model is applied to predict the parallel β -helix fold, an important fold in bacterial infection of plants and binding of antigens. The cross-family validation shows that SCRFs not only can score all known β -helices higher than non β -helices in Protein Data Bank, but also demonstrate more success in locating each rung in the known β -helix proteins than BetaWrap, a state-of-the-art algorithm for predicting β -helix fold, and HMMER, a general motif detection algorithm based on HMM. Applying our prediction model to Uniprot database, we hypothesize previously unknown β -helices.

1 Introduction

It is believed that protein structures reveal important information about the protein functions. One key step towards modeling a tertiary structure is to identify how secondary structures as building blocks 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 α - and $\beta\beta$ -hairpins [1, 2, 3, 4]. The task of protein fold recognition is the following: given a protein sequence and a particular fold or super-secondary structure, predict whether the protein contains the structural fold and if so, locate its exact positions in the sequence.

The traditional approach for protein fold prediction is to search the database using PSI-BLAST [5] or match against an HMM profile built from sequences with the same fold by HMMER [4] or SAM [3]. 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 folds in β class [6]. These cases necessitate a more powerful model, which can capture the structural characteristics of the protein fold. Interestingly, the protein fold recognition task parallels an emerging trend in machine learning community, i.e the *structure* prediction problem, which predict the labels of each node in a graph given the 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 [7, 8].

In fact, several graphical models have been applied to protein structure prediction. One of the early approaches is to apply simple hidden markov models (HMMs) to protein secondary structure prediction and protein motif detection [3, 4, 9]; Delcher et al. introduced probabilistic causal networks for protein secondary structure modeling [10]. Recently, Liu et al. applied conditional random fields (CRFs), a discriminative graphical model based on undirected graph, for protein secondary structure prediction [11]; Chu et al. extended segmental semi-Markov model (SSMM) under the Bayesian framework for protein secondary structures [12].

The bottleneck for protein fold prediction is the long-range interactions, which 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 generating process, which makes it difficult to consider overlapping features and long-range interactions. Discriminative graphical models, such as CRFs, assume a single residue as an observation. 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 graphical chain 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) [13] and Conditional Random fields (CRFs) [14]. Among these models, CRFs proposed by Lafferty et al., are very effective in many applications, including information extraction, image processing and so on [8, 7].

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 Hammersely-Clifford theorem [15], the conditional probability $P(\mathbf{y}|\mathbf{x})$ is proportional to the product of the potential functions over all the cliques in the graph,

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

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 [16] for more detail). For a chain structure, CRFs define the conditional probability as

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

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 conditional likelihood function is convex so that finding the global optimum is guaranteed [14]. Second, the feature function can be arbitrary, including overlapping features and long-range interactions. Finally, CRFs still have efficient algorithms, such as forward-backward or Viterbi, as long as the graph structures are sequences or trees.

Similar 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 is:

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

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:

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

The forward-backward and Viterbi algorithms can be derived accordingly [17].

3 Segmentation Conditional Random Fields (SCRFs)

Protein folds are frequent arrangement pattern of several secondary structure elements: some elements are quite conserved or prefer a specific length, while

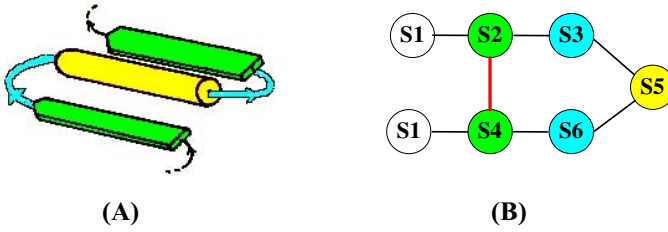


Fig. 1. Graph structure of β - α - β motif (A) 3-D structure (B) Protein structure graph: node: Green= β -strand, yellow= α -helix, cyan=coil, white=non- β - α - β (I-node); edge: $E_1 = \{\text{black edges}\}$ and $E_2 = \{\text{red edges}\}$

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 (or node) and the edges between elements indicating their interactions in 3-D. Then, given a protein sequence, we can search for the best segmentation defined by the graph and determine if the protein has the fold.

3.1 Protein Structural Graph

Before covering the algorithm in detail, we first introduce a special kind of graph, called protein structural graph. Given a protein fold, a structural graph is defined as $G = \langle V, E_1, E_2 \rangle$, where $V = U \cup \{I\}$, 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. 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: similar to HMMs, the E_1 edges indicate transitions of states 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.

In practice, one protein fold might correspond to several reasonable structural graphs given different semantics for one node. There is always a tradeoff between the graph complexity, fidelity of 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 less important or less interesting to us. We give a concrete example of β -helix fold in Section 4.

3.2 Segmentation Conditional Random Fields

Since a protein fold is regular arrangement of its secondary structure elements, the general topology is often known apriori 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. In the following discussion, we focus on this simplified but common case.

Consider the graph $G' = \langle V, E_2 \rangle$, given a protein sequence $\mathbf{x} = x_1x_2 \dots x_N$, we can have a possible segmentation of the sequence, i.e. $S = (S_1, S_2, \dots, S_M)$, 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 label of the segment y_i . The conditional probability of a segmentation S given the observation \mathbf{x} can be computed as follows:

$$P(S|\mathbf{x}) = \frac{1}{Z_0} \prod_{c \in G'(S, \mathbf{x})} \exp\left(\sum_k \lambda_k f_k(\mathbf{x}_c, S_c)\right),$$

where Z_0 is a normalization factor. If each subgraph of G' is a chain or a tree (an isolated node can also be seen as a chain), then we have

$$P(S|\mathbf{x}) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^M \sum_{k=1}^K \lambda_k f_k(\mathbf{x}, S_i, S'_{i-1})\right), \tag{2}$$

where S'_{i-1} is the direct forward neighbor of S_i in graph G' .

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

$$L_A = \sum_{i=1}^M \sum_{k=1}^K \lambda_k f_k(\mathbf{x}, S_i, S'_{i-1}) - \log Z_0 + \frac{\lambda_k^2}{2\sigma^2},$$

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

$$\frac{\partial L_A}{\partial \lambda_k} = \sum_{i=1}^M (f_k(\mathbf{x}, S_i, S'_{i-1}) - E_{P(s|\mathbf{x})}[f_k(\mathbf{x}, S_i, S'_{i-1})]) + \frac{\lambda_k}{\sigma^2}, \tag{3}$$

where $E_{P(s|\mathbf{x})}[f_k(x, S_i, S'_{i-1})]$ is the expectation of feature $f_k(\mathbf{x}, S_i, S'_{i-1})$ over all possible segmentations of x . The convexity property guarantees that the root corresponds to the optimal solution. However, since there is no closed-form solution to (3), it is not straightforward to find the optimal. 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 [17], which is also confirmed in our experiments.

Similar to CRFs, we still have an efficient inference algorithm as long as each subgraph of G' is a chain. 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 $x_{l+1} \dots x_r$ and a segment of state y_l ends at position l . The recursive step can be written as:

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

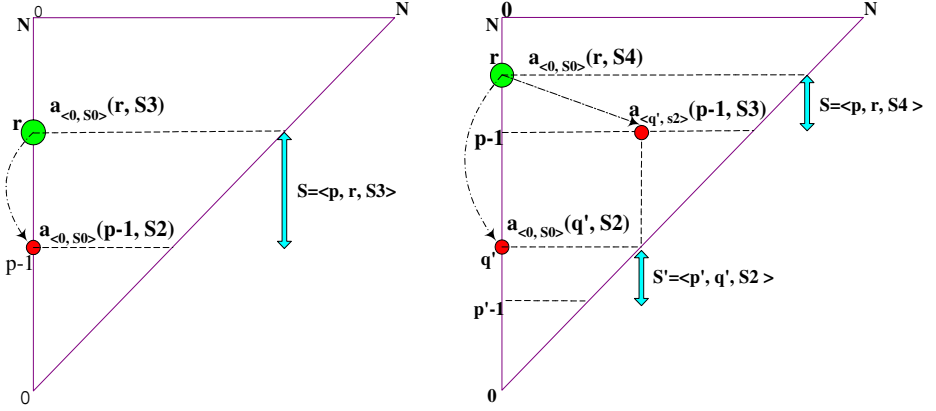


Fig. 2. An example of forward algorithm for the graph defined in Figure-1(B). x/y-axis: index of starting/end residue position; green circle: target value; red circle: intermediate value. (Left) calculation for $\alpha_{\langle 0, S_0 \rangle}(r, S_3)$ for segment S_3 with no direct forward neighbor;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

where S' is the direct forward neighbor of S in graph G' (if any), $S = \langle p, r, y_r \rangle$, $S' = \langle p', q', y' \rangle$, “ \rightarrow ” is the operator for next state and “ \leftarrow ” for previous state (the value is known since the state transition is deterministic). The range over the summation is $\sum_{p=r-\ell_2+1}^{r-\ell_2+1} \sum_{q'=l+\ell'_1-1}^{p-1} \sum_{p'=l}^{q'-\ell'_1+1}$, where $\ell_1 = \max \text{length}(y)$, $\ell_2 = \min \text{length}(y)$. Then the normalizer $Z_0 = \alpha_{\langle 0, y_{\text{start}} \rangle}(N, y_{\text{end}})$. 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 $x_{l+1} \dots x_r$ given a segment of state y_l ends at l and a segment of state y_r ends at r . Then we have

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

where $S = \langle p, q, y \rangle$, $S' = \langle l+1, q', \overleftarrow{y_l} \rangle$. 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 $P(S|x)$. Similar to CRFs, we define:

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

The best segmentation is the path traced back by $\max \delta_{\langle 0, y_{\text{start}} \rangle}(N, y_{\text{end}})$, 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. Therefore the final complexity will be approximately $O(N^2)$.

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 [18]. The typical 3-D structure of a β -helix is shown in Fig. 3(A-B). As we can see, each basic structural unit, i.e. a rung, has three β -strands of various lengths, ranging from 3 to 5 residues. The strands are connected to each other by loops with distinctive features. One loop is a unique two-residue turn which forms an angle of approximately 120° between two parallel β -strands (called *T-2 turn*). The other two loops vary in size and conformation, which might contain helix or even β -sheets. There currently exist 14 protein sequences with three-stranded right-hand β -helix whose crystal structures have been deposited in Protein Data Bank (PDB) (See Table 1). The β -helix structures are significant in that they include pectate lyases, which are secreted by pathogens and initiate bacterial infection of plants; the phage P22 tailspike adhesin that 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.

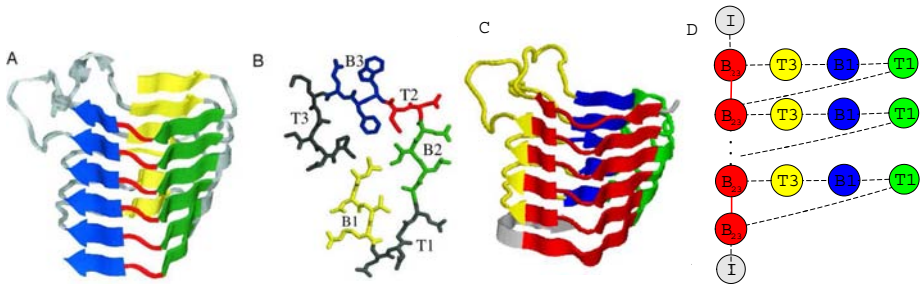


Fig. 3. 3-D structures and side-chain patterns of β -helices; (A) Side view (B) top view of one rung (C) Segmentation of 3-D structures (D) protein structural graph. E1 = {black edge} and E2 = {red edge}

Traditional methods for protein family classification, such as threading, PSI-BLAST and HMMs, fail to solve the β -helix recognition problem across different families [19]. Recently, a computational method called BetaWrap, has been proposed to predict the β -helix specifically [19]. The algorithm “wraps” the unknown sequences in all plausible ways and check the scores to see if any wrap makes sense. The cross-validation results in the protein data bank (PDB) seem promising. However, the BetaWrap algorithm suffers from hand-coding many biological heuristic rules. Hence it is prone to over-fit the known β -helix proteins and hard to generalize for other structural prediction tasks.

4.1 Protein Structural Graph for β -Helix

From previous literature on β -helix, 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; 2) the interaction of the strand side-chains in the buried core are critical determinants of the fold [20, 21]. Therefore we define the protein structural graph of β -helix as in Fig.3-(D).

There are 5 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 number 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 so that it varies from 1 to 80. s-I is the non- β -helix state, which refers to all those regions outside the β -helix structures. The red edge 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; b) each subgraph of $G' = \langle V, E_2 \rangle$ is a chain or a tree. For 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 that graph G' consists of a chain and a set of isolated nodes. Therefore the algorithm discussed in Section 3.2 can be applied accordingly.

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 state s-I, i.e. $\rho = \log \frac{\max_s P(S|x)}{P(\langle 1, N, s-I \rangle | x)}$. 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 3 residues in s-B1 segments. However, we use the B1 interactions as a filter in Viterbi algorithm: specifically, $\delta_t(y)$ will be the highest value whose corresponding segmentation also have alignment scores for B1 higher than some 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 *inter-node features*.

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

a) Regular expression template: Based on the side-chain alternating patterns in B23 region, BetaWrap generates a regular expression template to detect β -helices, i.e. $\Phi X \Phi X X \Psi X \Phi X$, where Φ matches any of the hydrophobic residues as {A, F, I, L, M, V, W, Y}, Ψ matches any amino acids except ionisable residues as {D, E, R, K} and X matches any amino acid [19]. Following similar idea, we define the feature function $f_{RST}(x, S)$ equal to 1 if the segment S matches the template, and 0 otherwise.

b) Probabilistic HMM profiles: The regular expression template as 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 [4] for the s-B23 and s-B1 segments respectively. We define the feature function $f_{HMM1}(x, S)$ and $f_{HMM2}(x, S)$ as the alignment scores of S 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 dimension. 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 function $f_{ssH}(x, S)$, $f_{ssE}(x, S)$ and $f_{ssC}(x, S)$ as the average of the predicted scores over all residues in segment S , for helix, sheet and coil respectively by PSIPRED [22].

d) Segment length: It is interesting to notice that the β -helix structure has strong preferences for insertions within certain length ranges. To consider this preference in the model, we did parametric density estimation. Several common functions are explored, including Poisson distribution, negative-binomial distribution and asymmetric exponential distribution, which consists for two exponential functions meeting at one point. We use the latter one since it provides a better estimator than the other two. Then we define the feature function $f_{L1}(x, S)$ and $f_{L3}(x, S)$ as the estimated probability of the length of segment S as s-T1 and s-T3 respectively.

Inter-node 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 of type X will align with residue Y, given their orientation relative to the core, is estimated from a β -structure database developed from the whole PDB [19]. Following similar idea, we define the feature function $f_{SAS}(x, S, S')$ as the weighted sum of the

side chain alignment scores for S given S' 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 is to study the different preferences for parallel and anti-parallel β -sheets. Steward & Thornton [23] 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

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

Compared with 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 $f_{PAS}(x, S, S') = score(S, S')$ if both S and S' are 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 get an good estimation of this distribution since the range is too large. Therefore we simply define the feature function as the normalized length, i.e. $f_{DIS}(x, S, S') = \frac{dis(S, S') - \mu}{\sigma}$, where μ is the mean and σ^2 is the variance.

It is interesting to notice that some features defined above are quite general, not limited to predicting β -helices only. For example, an important aspect to discriminate a specific protein fold with 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 certain length range, then length are also informative. For internode features, the β -sheet alignment scores are useful for folds in β -family while hydrophobicity is important for α - or $\alpha\beta$ -family.

5 Experiments

In our experiments, we followed the setup described in [19]. A PDB-minus dataset was constructed from the PDB protein sequences (July 2004 version) [24] 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 2094 sequences in total. The proteins in PDB-minus dataset will serve as negative examples in the cross-family validation and 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 [1]. For each cross, proteins in the

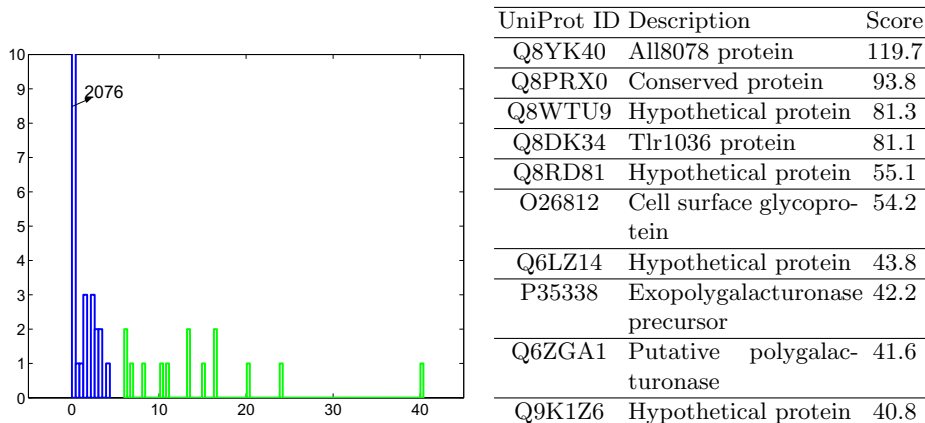
one β -helix family are placed in the test set while the remainder are placed in the training set as positive examples. Similarly, the PDB-minus was also randomly partitioned into nine subsets, one of which are placed in the test set while the rest serve as the negative training examples. We compare our results with BetaWrap, a 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 sequence-based method. Therefore we generated two kinds of alignments for comparison: one is the multiple structural alignments using EC-MC [25], the other is purely sequence-based alignments by CLUSTALW[26].

Table 1 shows the output scores by different methods and the relative rank for 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 PDB. On the other hand, there are two proteins (i.e. 1ktw and 1ea0) in our validation sets that are crystallized recently and thus are not included in the BetaWrap system. We test these two sequences on BetaWrap and get a score of -23.4 for 1ktw and -24.87 for 1ea0. These values are significantly

Table 1. Scores and rank for the known right-handed β -helices by HMMER, BetaWrap and SCRFs. 1: the scores and rank from BetaWrap are taken from [3] except 1ktw and 1ea0; 2: the bit scores in HMMER are not directly comparable

| SCOP family | PDB-id | Struct-based HMMs | | Seq-based HMMs | | BetaWrap ¹ | | SCRFs | |
|-----------------------|--------|------------------------|------|------------------------|------|-----------------------|------|---------------|------|
| | | Bit score ² | Rank | Bit score ² | Rank | Score | Rank | ρ -score | 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 |

Table 2. (Left) Histograms of protein scores of known β -helix proteins against PDB-minus dataset. Blue bar: PDB-minus dataset; green bar: known β -helix proteins. 2076 out of 2098 protein sequences in PDB-minus have a log ratio score ρ of 0, which means that the best segmentation is a single segment in non- β -helix state; (Right) Examples of proteins predicted to form β -helix in UniProt




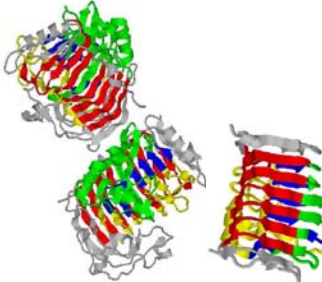
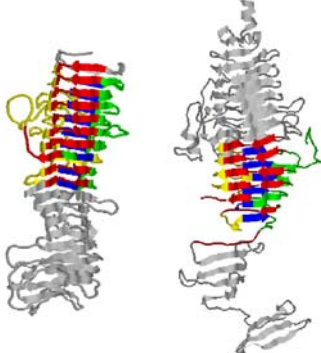
lower than the scores of other β -helices and some of the non β -helix proteins, which indicates that the BetaWrap might be overtrained. As expected, HMMER did worse than SCRFs and BetaWrap even using the structural alignments.

Table 2 plots the score histogram for known β -helix sequences against the PDB-minus dataset. Compared with the histograms in similar experiment by BetaWrap [19], our log ratio score ρ indicates a clearer separation of β -helix proteins v.s. non β -helix proteins. Only 18 out of 2094 proteins has a score higher than 0. Among these 18 proteins, 13 proteins belong to the beta class and 5 proteins belong to the alpha-beta class in CATH database [2]. In Table 3 we also cluster the proteins into three different groups according to the segmentation results and show examples of the predicted segmentation in each group.

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-Jul-2004 and TrEMBL 27.2 of 30-Jul-2004) [27]. We choose the UniRef50 (50% identity) with 490,713 sequences as the discovering set. 93 sequences were returned with scores above a cutoff of 5, 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. It is interesting to note that none of the known β -helices are from plants. The remaining eukaryotic sequences come from mammals, fungi, nematodes and pathogens from the genus Plasmodium: 4 sequences were viral, including 3 from bacteriophages; 9 sequences are archeal, 7 of which are from methanogens of the genus Methanosarcina. Of the 93 high scoring se-

Table 3. 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) |
| |  |  |  |

quences, 48 are likely homologous (BLAST E-val < 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 PDB. The protein sequences with maximal log ratio scores is shown in Table 2 (the full list can be accessed at <http://www.cs.cmu.edu/~yanliu/SCRF.html>).

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 tailspike 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. A 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 less sequence similarity.

6 Discussion and Conclusion

In [19], 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 ask: why is β -helix prediction difficult for these commonly used methods? why can SCRFs model perform better?

We think 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 β -helix are not reflected clearly in the sequences. For example, the conserved templates for 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 a β -helix, 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.

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. On the other hand, the 2-D protein structural graph has limited power to capture the dynamic constraints for 3-D protein structures. Therefore it would be interesting to extend the SCRFs model to include protein dynamics. The latter, however, will be a major undertaking.

Acknowledgement

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

1. Murzin, A., Brenner, S., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* **247** (1995) 536–40
2. Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., Thornton, J.: CATH—a hierarchic classification of protein domain structures. *Structure.* **5** (1997) 1093–108
3. Karplus, K., Barrett, C., Hughey, R.: Hidden markov models for detecting remote protein homologies. *Bioinformatics* **14** (1998) 846–56
4. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press (1998)
5. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped BLAST and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25** (1997) 3389–402
6. Menke, M., Scanlon, E., King, J., Berger, B., Cowen, L.: Wrap-and-pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils. In: *Proceedings of the 8th ACM RECOMB conference.* (2004) 298–307

7. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. In: Proc. IEEE International Conference on Computer Vision (ICCV). (2003) 1150–1159
8. Pinto, D., McCallum, A., Wei, X., Croft, W.B.: Table extraction using conditional random fields. In: Proceedings of the 26th ACM SIGIR conference. (2003) 235–242
9. Byströf, C., Thorsson, V., Baker, D.: HMMSTR: a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol.* **301** (2000) 173–90
10. Delcher, A., Kasif, S., Goldberg, H., Xsu, W.: Protein secondary-structure modeling with probabilistic networks. In: International Conference on Intelligent Systems and Molecular Biology (ISMB'93). (1993) 109–117
11. Liu, Y., Carbonell, J., Klein-Seetharaman, J., Gopalakrishnan, V.: Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics.* **20** (2004) 3099–107
12. W. Chu, Z.G., Wild, D.L.: A graphical model for protein secondary structure prediction. In: Proc. of International Conference on Machine Learning (ICML-04). (2004) 161–168
13. McCallum, A., Freitag, D., Pereira, F.C.N.: Maximum entropy markov models for information extraction and segmentation. In: Proc. of International Conference on Machine Learning (ICML-00). (2000) 591–598
14. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 282–289
15. Hammersley, J., Clifford, P.: Markov fields on finite graphs and lattices. Unpublished manuscript (1971)
16. Jordan, M.I.: *Learning in Graphical Models.* The MIT press (1998)
17. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of Human Language Technology, NAACL 2003. (2003)
18. Yoder, M., Keen, N., Jurnak, F.: New domain motif: the structure of pectate lyase c, a secreted plant virulence factor. *Science* **260** (1993) 1503–7
19. Bradley, P., Cowen, L., Menke, M., King, J., Berger, B.: Predicting the beta-helix fold from protein sequence data. In: Proceedings of 5th Annual ACM RECOMB conference. (2001) 59–67
20. Yoder, M., Jurnak, F.: Protein motifs. 3. the parallel beta helix and other coiled folds. *FASEB J.* **9** (1995) 335–42
21. Kreisberg, J., Betts, S., King, J.: Beta-helix core packing within the triple-stranded oligomerization domain of the p22 tailspike. *Protein Sci.* **9** (2000) 2338–43
22. Jones, D.: Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292** (1999) 195–202
23. Steward, R., Thornton, J.: Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins.* **48** (2002) 178–91
24. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. *Nucleic Acids Research* **28** (2000) 235–42
25. Guda, C., Lu, S., Sheeff, E., Bourne, P., Shindyalov, I.: CE-MC: A multiple protein structure alignment server. *Nucleic Acids Res.* **In press** (2004)
26. Thompson, J., Higgins, D., Gibson, T.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22** (1994) 4673–80
27. Leinonen, R., Diez, F., Binns, D., Fleischmann, W., Lopez, R., Apweiler, R.: Uniprot archive. *Bioinformatics.* **20** (2004) 3236–7