Conditional Graphical Models for Protein Structural Motif Recognition

Citation	Liu, Yan et al. "Conditional Graphical Models for Protein Structural Motif Recognition." Journal of Computational Biology 16.5 (2009) : 639-657. © 2009 Mary Ann Liebert, Inc.
As Published	http://dx.doi.org/10.1089/cmb.2008.0176
Publisher	Mary Ann Liebert, Inc.
Version	Final published version
Citable link	http://hdl.handle.net/1721.1/62177
Terms of Use	Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 5, 2009
© Mary Ann Liebert, Inc.
Pp. 639–657
DOI: 10.1089/cmb.2008.0176

Conditional Graphical Models for Protein Structural Motif Recognition

YAN LIU,[1] JAIME CARBONELL,[2] VANATHI GOPALAKRISHNAN,[3] and PETER WEIGELE[4]

ABSTRACT

Determining protein structures is crucial to understanding the mechanisms of infection and designing drugs. However, the elucidation of protein folds by crystallographic experiments can be a bottleneck in the development process. In this article, we present a probabilistic graphical model framework, conditional graphical models, for predicting protein structural motifs. It represents the structure characteristics of a structural motif using a graph, where the nodes denote the secondary structure elements, and the edges indicate the side-chain interactions between the components either within one protein chain or between chains. Then the model defines the optimal segmentation of a protein sequence against the graph by maximizing its "conditional" probability so that it can take advantages of the discriminative training approach. Efficient approximate inference algorithms using reversible jump Markov Chain Monte Carlo (MCMC) algorithm are developed to handle the resulting complex graphical models. We test our algorithm on four important structural motifs, and our method outperforms other state-of-art algorithms for motif recognition. We also hypothesize potential membership proteins of target folds from Swiss-Prot, which further supports the evolutionary hypothesis about viral folds.

Key words: conditional random fields, graphical models, protein structure prediction.

1. INTRODUCTION

Three-dimensional protein structures play key roles in determining the functions, activities, and subcellular localizations of proteins. An important step in automatically inferring protein structures from amino-acid sequences is to identify the typical spatial arrangements of well-defined secondary structures, which are conserved over proteins in different organisms and/or from different species, i.e., *structural motif*.

From computational perspective, we can represent a structural motif abstractly by a state sequence of its secondary structure components (such as α-helix, β-sheet, and coil) and the constraints between them (i.e., chemical bonding among secondary structure components). For example, one important motif in binding the

[1] IBM T.J. Watson Research Center, Yorktown Heights, New York.
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.
[3] Center for Biomedical Informatics, University of Pittsburgh, Pennsylvania.
[4] Biology Department, Massachusetts Institute of Technology, Cambridge, Massachusetts.

ligands or substrates is the β-α-β motif. It consists of three secondary structure components, including a β-strand, an α-helix, and another β-strand that forms hydrogen bonds with the first β-strand. Therefore, we can abstract the motif as a state sequence "B_1–A–B_2" and hydrogen bonding between the two β-strands as the constraints in the objective function. The goal of structural motif recognition is to predict whether the motif of interest exists in the testing protein sequence. Computationally, it can be achieved by segmenting and labeling the testing sequence against the motif template, i.e., the state sequence and the constraints.

From the machine learning perspective, protein structural motif recognition can be cast as a structured prediction problem. Structured prediction refers to the task in which the observed data are sequential or with other simple structures while the outputs actually involve complex structures. More specifically, in the task of protein structure prediction, we are given the observation of a sequence of amino acids, but the target outputs involves complex three-dimensional protein structures. By considering the constraints or associations between the outputs (beyond the i.i.d. assumption), we are able to achieve a better prediction performance.

Conditional graphical models, such as conditional random fields (CRF) (Lafferty et al., 2001), max-margin Markov networks (Taskar et al., 2003), semi-Markov CRF (Sarawagi and Cohen, 2004), and so on, have demonstrated successes to solve such problems in multiple applications. Therefore, we follow the graphical model approach and propose a series of new models for our task of structural motif recognition. These models can be seen as an extension of the CRF (Lafferty et al., 2001) by joint modeling the constraints between the components either on one sequence or multiple sequences. The key questions we address in this article are as follows: How can we better represent the structural information of a motif using the graphical model? Given the foreseeable complexity of the model, how can we learn the parameters of the model and make inferences efficiently?

The rest of the article is organized as follows: In Section 2 we introduce the basic concept in protein structures. Next, in Section 3, we define the conditional graphical models and demonstrate how to derive corresponding models for specific structural motifs with three examples. In the following two sections, Sections 4 and 5, we discuss the inference and learning algorithms as well as feature definitions. In Section 6, we show the experiment results. Finally, in Section 7, we conclude with a discussion and suggestions for future work.

2. PROTEIN STRUCTURAL MOTIF

Most of the essential structures and functions of the cells are realized by proteins, which are a chain of amino acids with stable three-dimensional structures. A fundamental principle in all of protein science is that protein functions are determined by their structures. However, it is extremely difficult to experimentally solve the structures of proteins. Therefore, how to predict protein structures from sequences using computational methods remains one of the most fundamental problems in structural bioinformatics and has been extensive studied for decades (Bourne and Weissig, 2003; Venclovas et al., 2003).

Protein structural motifs (or sometimes referred as protein folds) are identifiable spatial arrangements of secondary structures. It is observed that there exist only a limited number of topologically distinct folds in nature (around 1,000), although we have discovered millions of protein sequences. As a result, proteins with the same structural motifs often do not demonstrate sequence similarities. Uncovering the relationships between sequence and structures might reveal important evolutionary information.

To date there has been significant progress in the general structural motif recognition and alignment task, ranging from sequence similarity matching approaches (Altschul et al., 1997; Bateman et al., 2004; Durbin et al., 1998; von Ohsen et al., 2003), to threading algorithms based on physical forces or multimeric threading (Aloy and Russell, 2003; Fischer, 2000; Jones et al., 1992; Kelley et al., 2000; Lu et al., 2002; Rooman and Wodak, 1995), and to machine learning methods (Cheng and Baldi, 2006; Ding and Dubchak, 2000; Do et al., 2006; Kamisetty and Langmead, 2007; Sander et al., 2006). In addition, there are various studies on designing specialized algorithms for well-defined structural motifs or functional units, such as $\alpha\alpha$- and $\beta\beta$-hairpins (Durbin et al., 1998; Karplus et al., 1998; Murzin et al., 1995; Orengo et al., 1997), α-helical membrane proteins (Fleishman and Ben-Tal, 2006), as well as several complex folds, for instance β-helix (Bradley et al., 2001) and beta trefoils (Menke et al., 2004). Unfortunately there has been very limited work to focus on "under-representative" protein structural motifs, i.e., those motifs with unclear sequence similarity (under 25%), few positive examples in Protein Data Bank (PDB) (Berman et al., 2000), and usually exhibiting long range interactions within or between the polypeptide chains. An example is the triple β-spiral (TBS) fold, a processive homotrimer which serves as a fibrous connector from the main virus capsid to a

C-terminal knob that binds to host cell-surface receptor proteins. The fold has been identified to commonly exist in adenovirus (a DNA virus which infects both humans and animals), reovirus (an RNA virus which infects human) and bacteriophage PRD1 (a DNA virus infecting bacteria). However, the similarities between these protein sequences are very low (below 25% in sequence identity). Identifying more examples of these under-representative motifs not only will help biologists to confirm the hypothesis that it is a common fold in nature, but also may reveal important evolutionary relationships between the viral proteins. These special characteristics render previous methods inadequate for modeling those under-representative structural motifs, which motivates us to seek more sophisticated approaches to solve the problem.

The problem setting is as follows: given a target protein motif, as well as a set of N training sequences $x^{(1)}$, $x^{(2)}$,…, $x^{(N)}$, including both positive and negative examples with structural annotation, i.e., three-dimensional coordinates of each atom in the proteins, we want to predict whether a new test sequence x^{test} (without structural annotation) has the motif in its structure or not; and if yes, identify its specific location in the sequence. Here, the protein $x^{(i)}$ is a sequence of amino acids, represented by capital letters corresponding to 20 different types of amino acids. The information about the target motif is provided by the literature and domain experts, mostly on what are the structural components of the motif, how they form chemical bonds between each other, and if possible which bonds are most essential to maintain a stable structure. Therefore, one of our major tasks is to convert the descriptive domain information into a mathematical formulation and solve the problem effectively.

3. CONDITIONAL GRAPHICAL MODELS

Remember that our application starts from a target motif F that the biologists are interested in. The motif F can be simple, with only a few secondary structure elements (called "supersecondary structure"), or complex, with many structural components forming sophisticated bonding patterns (called "protein fold"). All the proteins with resolved structures deposited in the centralized database—namely, PDB (Berman et al., 2000)—can be classified into two groups: those taking the target motif F (i.e., positive examples), and those not (i.e., negative examples). These proteins together with the labels and structure annotations can be used as the training data. Our goal is to predict whether a testing protein sequence, without resolved structures, takes the motif F in nature or not; if they do, locate the starting and ending positions of the subsequence that takes the motif.

As we can see, the task involves two sub-tasks: one is the classification problem, that is, given a set of training sequences X_1,X_2, …, X_N and their labels y_1,y_2, …, y_N ($y_i = 0,1$), predict the label of a new testing sequence X_{new}; the other subtask is not straightforward to describe in mathematical settings, but we can think of the target fold as some patterns (or motifs in bioinformatics terminology). Given a set of instances of the pattern, including both the positive examples (subsequences with the pattern F) and the negative examples (sequences without the pattern F), we want to predict *where* the pattern appears in the testing protein sequences. The first question can be answered easily if we can solve the second one successfully. A key problem in the second task is how we can represent the descriptive patterns (or motifs) using mathematical notations. In structural biology, the conventional representation of a protein motif is a graph (Westhead et al., 1999), in which the nodes represent the secondary structure components and the edges indicate the inter- and intra-chain interactions between the components in the three-dimensional structures. This intuitive representation motivates us to apply graphical models, which combine graph theory and probability theory, for structural motif recognition. Next, we describe our graphical model approach to solve the problem.

3.1. Model definition

Given one structural motif we are interested in, an undirected graph $G = <\mathcal{V}, \mathcal{E}>$ can be constructed, where $\mathcal{V} = \mathcal{U} \bigcup \mathcal{I}$, $\mathcal{U}$ is the set of nodes corresponding to the secondary structure components inside the motif and I is the node to represent the component outside the motif. $\mathcal{E}$ is the set of edges between neighboring nodes denoting the chemical bonding between the components, either in linear sequence order (i.e., the polypeptide bonding) or in three-dimensional structures (i.e., chemical bonding, such as hydrogen bonds or disulfide bonds).

Figure 1 shows an example of the β-α-β motif we discussed earlier. Notice that each node corresponds to one structural component composed of multiple number of amino acids and the challenge is that the length is not fixed. Therefore, we need to infer the location of each node, i.e., segmenting the protein sequence against the graph. Starting with a structure graph G defined on one chain and an observed protein sequence $\mathbf{x} = x_1 x_2 \ldots x_N$, the random variables corresponding to the nodes in graph are: $\mathbf{Y} = \{M, \mathbf{W}\}$, where M denotes the number of nodes in the graph. Notice that M can be either a constant or a variable taking values from a discrete sets $\{1,\ldots, m_{\max}{}^1\}$ (depending on whether the target motif F has fixed number of structural components or not). $W = \{W_1,\ldots, W_M\}$ and $W_i = \{s_i, d_i\}$ is the label for the i^{th} node, specifically s_i is the state and d_i is the ending position. W_i completely determines the i^{th} node according to its semantics defined in the graph. Under this setup, a value instantiation of $\mathbf{Y}$ defines a unique segmentation and annotation of the observed protein sequence $\mathbf{x}$.

Getting a reasonable graph definition for a concerned motif requires domain knowledge and expertise. To make our discussion more focused, we assume that the following information is given: the graph G, the state set S, and a set of training sequences $\mathbf{x}$ with corresponding labels $\mathbf{y}$. Our goal for the rest of the section is to define a probabilistic framework to model the structural properties of the target motif and predict the segmentation for testing sequences.

A probabilistic distribution on a protein structural graph can be postulated using the potential functions defined on the *cliques* of nodes induced by the edges in the graph (Hammersley and Clifford, 1971). Discriminative models, such as conditional random fields (Lafferty et al., 2001), estimate the decision boundary directly without computing the underlying data distribution and thus often achieve better performance. Therefore, following the idea, we directly define the conditional probability of $\mathbf{Y}$ given the observation $\mathbf{x}$ as

$$P(\mathbf{Y}|\mathbf{x}) = \frac{1}{Z} \prod_{c \in CG} \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, \mathbf{Y}c)), \qquad (1)$$

where f_k is the k^{th} feature defined over the cliques c, such as the secondary structure assignment or the segment length; λ_k, the weight of the feature f_k, is the model parameters and has to be learned through the training data; Z is the normalization constant, namely $Z = \sum_{M=1}^{m^{max}} \sum_{d_0=1}^{L} \sum_{d_1=d_0+1}^{L} \cdots \sum_{d_M=d_{M-1}+1}^{L} \Pi_{c \in C_G} \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, \mathbf{Y}_c))$; C_G is the clique set of graph G. Since C_G can be a huge set, each Y_c can include a large number of nodes due to various levels of dependencies. Thus, designing features for such cliques is non-trivial because one has to consider all the joint configurations of all the nodes in a clique.

We described a general definition for the conditional graphical models above. In the following sections, we show a series of examples and explain how to define specific models based on the structural properties of target motif F: (1) fixed template motif; (2) repetitive motif; and (3) quaternary motif.

3.2. *Fixed template motif*

By definition, structural motifs are regular arrangement of secondary structures. Therefore, the spatial ordering of most protein motifs is fixed, i.e., the structural components and how they are connected to each other are predetermined (therefore, the nodes and the edges in G are fixed), which leads to a deterministic dependency between nodes. The β-α-β motif is an example of the fixed template motifs. Their structural properties result in a simplification of the "effective" clique sets (those need to be parameterized) and the relevant feature design. In other words, we can define a set of states $S = \{s_1, \ldots, s_M\}$, with each state corresponding to one node. Such definitions lead to deterministic state transition, i.e., $P(s_i | s_{i\text{-}1}) = 1$. For a testing sequence, we only need to infer the starting position of each node, i.e., $\{d_i\}$. Therefore only pairs of non-local cliques, e.g., those connected by the undirected "red" arc in Figure 1, need to be modeled. By considering the pairwise edge potentials, we can achieve the following simplified formulation:

$$P(\mathbf{Y}|\mathbf{x}) = P(\{d_i\}|\mathbf{x}) = \frac{1}{Z} \exp(\sum_{i=1}^{M} \sum_{k=1}^{K_1} \lambda_k f_k(\mathbf{x}, d_{i-1}, d_i)) \exp(\sum_{\{i,j\} \in \mathcal{E}'} \sum_{k=1}^{K_2} \lambda_k g_k(x, d_{i-1}, d_i, d_{j-1}, d_j)), \qquad (2)$$

where ε' denotes the set of non-local edges; the feature f_k is defined over the i^{th} node (i.e., subsequences $x_{di\text{-}1+1} \ldots x_{d_i}$); the feature g_k is defined over the pair of nodes (i.e., subsequences $x_{d_{i\text{-}1}+1} \ldots x_{di}$ and $x_{dj\text{-}1+1} \ldots x_{dj}$).

[1] $m_{\max}$ is the maximal number of nodes allowed (usually defined by the biologists).

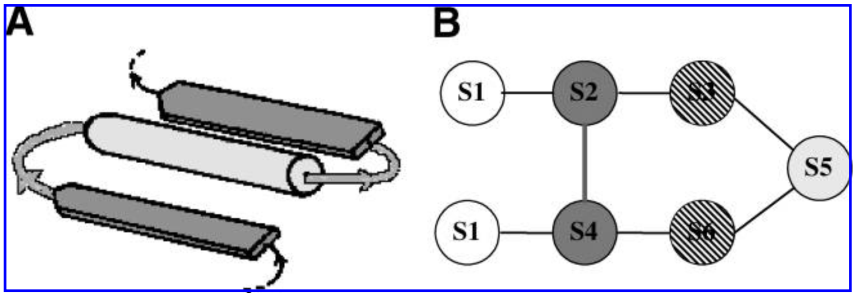


FIG. 1. Graph structure of β-α-β motif (**A**) and 3-D structure (**B**) Graphical model representation: node: Dark grey = β-strand; light grey = α-helix; white = non-β-α-β (I-node); edge: thin black = local edges and thick black = non-local edges.

3.3. Repetitive motif

Previous approaches in computational biology, such as sequence-based methods and hidden Markov model (HMM)–like methods, perform reasonably well on the fixed template motifs. However, they fail in accurately predicting more complex and irregular protein motifs as those containing highly stochastic (in terms of sequence composition, spacing, and ordering) internal structures, for example, the repetitive motifs and quaternary motifs.

Repetitive motifs are defined as a variable number of repeats of a fixed template motif. One example of the repetitive motifs is the right-handed parallel β-helix. It is an elongated helix-like structure with a series of progressive stranded coilings (repeats), each of which is composed of three parallel β-strands to form a triangular prism shape (Yoder et al., 1993). The typical three-dimensional structure of a β-helix is shown in Figure 2A, B. As we can see, each basic structural unit, i.e., a repeat, 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. The repetitive motifs are believed to be prevalent in many proteins and involve in a wide spectrum of cellular and biochemical activities, such as the initiation of bacterial infection (Yoder et al., 1993) and various protein-protein interaction processes (Kobe and Deisenhofer, 1994). The major challenges in computationally predicting these motifs are as follows: (1) the long-range interactions between their build-blocks (i.e., structural motifs) due to unknown number of spacers (i.e., amino acid insertions) between adjacent motifs; and (2) low sequence similarities between those motif repeats within the same protein and also across proteins.

We consider the corresponding conditional graphical models for the repetitive motifs. For one motif repeat, we can construct a state set similar as the fixed template motifs. Since all the motif instances within one protein or across proteins have the same structure properties, we will share the state sets for all the repeats. Different from the fixed template motifs, we do not know the number of nodes beforehand since the

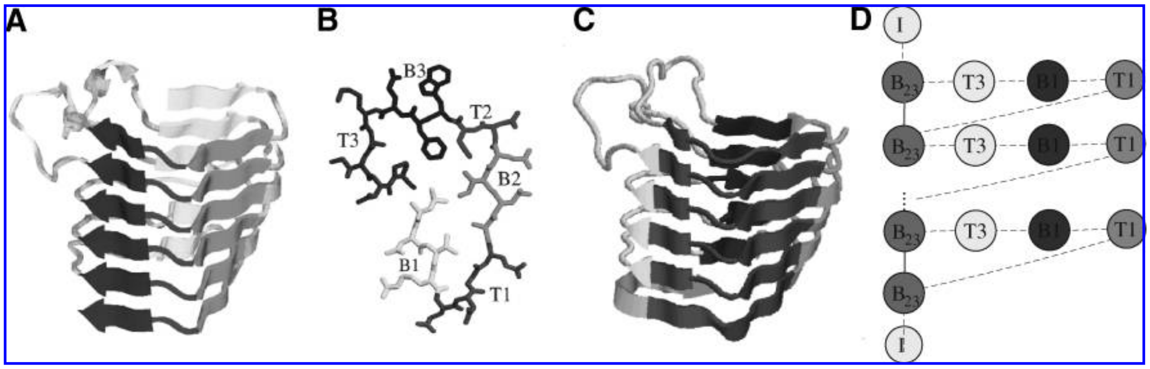


FIG. 2. 3-D structures and side-chain patterns of β-helices. (**A**) Side view. (**B**) Top view of one rung. (**C**) Segmentation of 3-D structures. (**D**) Graphical model representation. Local edges (black) and non-local edges. A and B are adapted from [9].

number of repeats differ from proteins to proteins. Therefore, M need to be inferred, but given a value of M, the states of each node is deterministic. Thus, we need only to infer M and $\{d_i\}$, that is,

$$P(\mathbf{Y}|\mathbf{x}) = P(M, \{d_i\}|\mathbf{x}) = P(M|\mathbf{x})P(\{d_i\}|\mathbf{x}, M). \tag{3}$$

Assuming $P(M|\mathbf{x})$ takes a uniform distribution, we have

$$P(\mathbf{Y}|\mathbf{x}) \propto \frac{1}{Z}\exp(\sum_{i=1}^{M}\sum_{k=1}^{K_1}\lambda_k f_k(\mathbf{x}, d_{i-1}, d_i))\exp(\sum_{\{i,j\}\in\mathcal{E}'}\sum_{k=1}^{K_2}\lambda_k g_k(\mathbf{x}, d_{i-1}, d_i, d_{j-1}, d_j)). \tag{4}$$

Note that the normalization constant Z in eq. (4) has to sum over all possible values of M as well as the corresponding set of d_i.

3.4. Quaternary motif

Quaternary motifs consist of *multiple* protein chains that form chemical bonds among the side chains of sequence-distant residues to reach a structurally stable domain. These motifs play very important roles in protein functions, such as enzymes, hemoglobin, DNA polymerase, and ion channels. One example of the quaternary motifs is the TBS (Fig. 3). It is a processive homotrimer consisting of three identical interacting protein chains with a series of repeated structural elements, each of which is composed of a β-strand, a long solvent-exposed loop, a second β-strand that forms antiparallel β-sheets with the first one but on a different protein chain, and a tight β-turn (Scanlon, 2004; van Raaij et al., 1999; Weigele et al., 2003). The fold serves as a fibrous connector from the main virus capsid to a C-terminal knob that binds to host cell-surface receptor proteins. Up to now, there are only three identified examples of the TBS motif, but they are found in both the DNA viruses and RNA viruses. By identifying more TBS examples, we might be able to reveal important evolution relationships among the viral proteins and help drug design. The major challenges for predicting the quaternary motifs are as follows: (1) much fewer positive examples for training (the size of the quaternary motifs makes it difficult to resolve their structures via lab experiments); and (2) less sequence conservation (the functional sites on the quaternary structures are more apt to change in order to adapt to the environment).

Now we consider the conditional graphical models for quaternary motifs. For each protein chain, we can construct its graphical model following the discussion before. Then for the chemical bonding between the structural components on different chains, we can draw an edge between the corresponding nodes to represent the interactions (for its graphical model representation, see Fig. 3C). In this way, we can generalize the conditional graphical models to the more complex quaternary motifs as follows: given a set of protein sequences $\mathbf{x}_{(1)}$, …, $\mathbf{x}_{(C)}$, we have a segmentation initiation of each chain according to the graphical models defined for the target motif, i.e., $\{\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})\}$, where $M_{(i)}$ and $\mathbf{w}_{(i)}$ follows the same definition as before for the i^{th} chain. Similar to the tertiary motifs, there also exist some quaternary motifs with structural repeats and we do no know the number of repeats for the testing protein beforehand. Therefore, the number of nodes M is unknown and need to be inferred. Following the previous formulation, we have

$$\begin{aligned}
P(\mathbf{Y}|\mathbf{x}) &= P(\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(C)}|\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(C)}) \\
&= \frac{1}{Z}\prod_{\mathbf{y}_j^{(i)}\in\nu_G}\Phi(\mathbf{x}_{(i)}, \mathbf{y}_{(i),j})\prod_{\langle\mathbf{y}_{(i),j}, \mathbf{y}_{(p),q}\rangle\in\mathcal{E}_G}\Phi(\mathbf{x}_{(i)}, \mathbf{x}_{(p)}, \mathbf{y}_{(i),j}, \mathbf{y}_{(p),q}) \\
&= \frac{1}{Z}\exp(\sum_{i=1}^{C}\sum_{i=j}^{M}\sum_{k=1}^{K_1}\lambda_k f_k(\mathbf{x}_{(i)}, d_{(i),j-1}, d_{(i),j}))\exp(\sum_{\{(i,j),(p,q)\}\in\mathcal{E}'}\sum_{k=1}^{K_2}\lambda_k g_k(\mathbf{x}, d_{(i),j-1}, d_{(i),j}, d_{(p),q-1}, d_{(p),q}))
\end{aligned}$$

where Z is the normalizer over all possible segmentation assignments of *all* component sequences. Notice that the joint modeling of all the component sequences are essential since the chemical bonding between the structural components on different chains directly determine the stability of the quaternary motifs.

By now we have examined three examples to demonstrate how to define reasonable conditional graphical models for different types of structural motifs. These examples represent a large population of the currently known motifs. As we can see, the conditional graphical models are general enough to model even the most complex structural motifs.

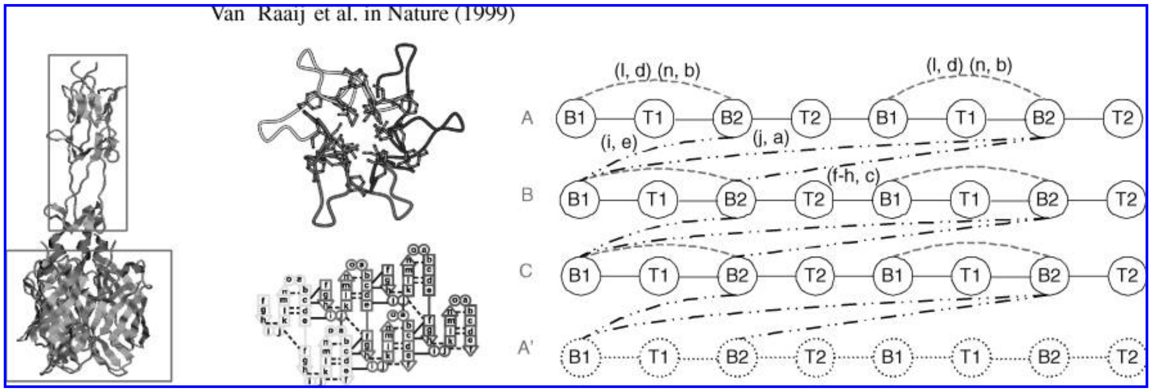


FIG. 3. (**Left**) Demonstration graph of triple β-spirals: 3-D structures view. First block, shaft region (target fold); second block, knob region. (**Middle**) Top view and maps of hydrogen bonds within a chain and between chains. (**Right**) PSG of the Triple β-spirals. Chain C' is a mirror of chain C for better visual effects. Dotted line, inter-chain interactions; solid line, intra-chain interactions. The pairs of characters on the edge indicate the hydrogen bonding between the residues denoted by the characters.

4. LEARNING AND INFERENCE

In conditional graphical models, given an observation sequence $\mathbf{x} = x_1 x_2 \ldots x_N$, the *conditional* probability of a possible segmentation $\mathbf{Y} = \{M, \{W_i\}\}$ against the protein structure graph G, is defined as

$$P(\mathbf{Y}|\mathbf{x}) = \frac{1}{Z} \prod_{c \in C_G} \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, \mathbf{Y}_c)), \tag{5}$$

Like the CRF model, the parameters $\lambda = (\lambda_1, \ldots, \lambda_K)$ can be computed by minimizing the regularized log-loss of the training data $\mathcal{L}$, i.e.,

$$\lambda = \text{argmax } \mathcal{L}(\lambda) = \text{argmax} \left\{ \sum_{j=1}^{L} \log P(\mathbf{y}^{(j)}|\mathbf{x}^{(j)}) + \Omega(\|\lambda\|) \right\}, \tag{6}$$

where L is the number of training sequences. The conditional likelihood function is convex so that finding the global optimum is guaranteed. Since there is no closed form solution to the optimization function above, we compute the first derivative of right side of eq. (6) with respect to λ and set it to zero, resulting in the equation below:

$$\sum_{j=1}^{L} f_k(\mathbf{x}^{(j)}, \mathbf{y}_c^{(j)}) - \sum_{j=1}^{L} E_{P(\mathbf{Y}|\mathbf{x}^{(j)})}[f_k(\mathbf{x}^{(j), \mathbf{Y}_c})] + \Delta\Omega(\|\lambda\|) = 0 \tag{7}$$

The intuition of eq. (7) is to seek the direction of λ_k where the model expectation agrees with the empirical distribution.

Given a testing sequence, our goal is to seek the segmentation configuration with the highest conditional probability defined above, i.e.,

$$\mathbf{Y}^{\text{opt}} = \text{argmax} \sum_{c \in C_G} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, \mathbf{Y}_c). \tag{8}$$

It can be seen that we need to compute the expectation of the features over the models in eq. (7) and search over all possible assignments of the segmentation to ensure the maximum in eq. (8). It is known that the complexity of the inference algorithm depends on the graphs in the models. If it is a simple chain, or a tree-like structure, we can use exact inference algorithms, such as belief propagation. For complex graphs, computing exact marginal distributions is in general infeasible and approximation algorithms have to be applied. In addition, there are millions of sequences in the protein sequence database. Such large-scale applications demand efficient inference and optimization algorithms. Therefore, a naive exhaustive search would be prohibitively expensive due to the complex graphs induced by the protein structures. In this section, we discuss a general inference and learning approach, which is able to handle the corresponding conditional graphical models efficiently for different types of structural motifs.

4.1. Training phase

In the training phase, we need to learn the model parameters λ by solving eq. (7). There is no closed form solution, therefore an iterative searching algorithm has to be applied. Recent advance on iterative searching algorithms suggests that the Langevin methods converge much faster than other commonly used methods, such as iterative scaling or conjugate gradient (Murray and Ghahramani, 2004; Yang et al., 2007). Therefore, we apply the Langevin methods to learn the model parameters.

Iterative Searching Using Langevin Monte Carlo. The *uncorrected Langevin* method originates from the Langevin Monte Carlo method by accepting all the proposed moves (Murray and Ghahramani, 2004). It makes use of the gradient information and resembles noisy steepest descent. The uncorrected Langevin form is expressed as follows:

$$\lambda_k^{\text{new}} = \lambda_k + \frac{\epsilon^2}{2}\frac{\partial}{\partial \lambda_k}\mathcal{L}(\lambda) + \epsilon n_k$$

where $n_k \sim N(0, 1)$. Intuitively, this rule performs gradient descent but explores a neighborhood around the optimum through the noise term. Taking the first derivative of the log likelihood $\mathcal{L}(\lambda)$, we have

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = \sum_{j=1}^{L}\{\boxed{f_k(\mathbf{x}^{(j)}, \mathbf{y}_c^{(j)}) - E_{P(\mathbf{Y}|\mathbf{x}^{(j)})}[f_k(\mathbf{x}^{(j)}, \mathbf{Y}_c)]}\} + \frac{\lambda_k}{\sigma^2} \tag{9}$$

As discussed before, the graphical models for predicting the structural motif are usually a complex graph with loops and multiple chains. Therefore, we need efficient approximation methods to estimate the terms inside the box on the right-hand side of eq. (9), which are referred to as $\triangledown\lambda_k$ later in our discussion.

Approximate Inference Using Contrastive Divergence. There are three major approximation approaches in graphical models: sampling, variational methods, and loopy belief propagation. Sampling techniques have been widely used in the statistics community; however; there are two main problems: inefficiency due to the long "burn-in" periods and large variance in the final estimation. To avoid these problems, we use contrastive divergence (CD), as proposed in Welling and Hinton (2002). CD is similar to Gibbs sampling, except that, instead of running Gibbs sampling until the equilibrium distribution is reached, it runs the sampler up to only a few iterations and uses the resulting distribution to approximate the true model distribution. The algorithm is described in Algorithm 1.

There will be a problem if we use the naive Gibbs sampling in step (2), since the segmentation hidden variables $\mathbf{w}_i$ may be of different dimensions in each sampling iteration, depending on the value of M (the number of structural components in the i^{th} sequence). The reversible jump Markov chain Monte Carlo (MCMC) algorithm has been proposed to solve the problem, with the ability to even handle the observations of multiple sequences as in quaternary structural motifs. It has demonstrated successes in various applications, such as mixture models, HMM for DNA sequence segmentation (Boys and Henderson, 2001), and phylogenetic trees (Huelsenbeck et al., 2004).

Reversible Jump Markov Chain Monte Carlo. We use the example of predicting quaternary structural motif (as discussed in Section 3.4) to demonstrate how to use reversible jump MCMC for

Algorithm 1 Description of contrastive divergence

```
      Input: $\lambda$; Output: $\triangledown\lambda$
1.    Sample a data vector $\mathbf{y}^0$ from the empirical distribution $P^0$;
2.    Iterate over T times:
          Sample a value for each latent variable $\mathbf{y} = \{M,\{\mathbf{w}_i\}\}$ from its posterior
            probability defined in eq(5). The value is represented as $\hat{\mathbf{y}}^T$.
4.    Calculate the contrastive divergence as $\nabla\lambda = E_{\mathbf{y}^0}[f_k] - E_{\hat{\mathbf{y}}}[f_k]$.
```

inferences. Given a set of protein sequences $\mathbf{x}_{(1)}$, …, $\mathbf{x}_{(C)}$ and one segmentation initiation of each sequence $\{\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})\}$, our goal is propose a new move $\mathbf{y}^*_{(i)}$. To satisfy the detailed balance defined by the MCMC algorithm, auxiliary random variables v and v^* have to be introduced. The definitions for v and v^* should guarantee the *dimension-matching requirement*, i.e., $\dim(y_i) + \dim(v) = \dim(y^*_i) + \dim(v')$ and there is a one-to-one mapping from (y_i, v) to (y^*_i, v'), i.e., there exists a function Ψ so that $\Psi(y_i, v) = (y^*_i, v')$ and $\Psi^{-1}(y^*_i, v') = (y_i, v)$. Then the acceptance rate for the proposed transition from $\mathbf{y}_{(i)}$ to $\mathbf{y}^*_{(i)}$ is

$$\min\{1, \text{posterior ratio} \times \text{proposal ratio} \times \text{Jacobian}\}$$
$$= \min\left\{1, \frac{P(\mathbf{y}_{(1)}, \ldots, \mathbf{y}^*_{(i)}, \ldots, \mathbf{y}_{(C)} | \{\mathbf{x}_{(i)}\}) P(v')}{P(\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(i)}, \ldots, \mathbf{y}_{(C)} | \{\mathbf{x}_{(i)}\}) P(v)} \left| \frac{\partial(\mathbf{y}^*_i, v')}{\partial(\mathbf{y}_i, v)} \right| \right\},$$

where the last term is the determinant of the Jacobian matrix.

To construct a Markov chain on the sequence of segmentations, we define four types of Metropolis operators (Green, 1995):

1. *State switching:* Given a segmentation $\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})$, select a segment j uniformly from $[1, M]$, and a state value s' uniformly from state set S. Set $\mathbf{y}^*_i = \mathbf{y}_{(i)}$ except that $s^*_{i,j} = s'$.
2. *Position switching:* Given a segmentation $\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})$, select the segment j uniformly from $[1, M]$ and a position assignment $d' \sim \mathrm{U}[d_{(i),j-1} + 1, d_{(i),j+1} - 1]$. Set $\mathbf{y}^*_i = \mathbf{y}_{(i)}$ except that $d^*_{(i),j} = d'$.
3. *Segment split:* Given a segmentation $\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})$, propose $\mathbf{y}^*_i = (M^*_{(i)}, \mathbf{w}^*_{(i)})$ with $M^*_{(i)} = M_{(i)} + 1$ segments by splitting the j^{th} segment, where j is randomly sampled from $\mathrm{U}[1, M]$. Set $\mathbf{w}^*_{i,k} = \mathbf{w}_{i,k}$ for $k = 1, \ldots, j - 1$, and $\mathbf{w}^*_{(i),k+1} = \mathbf{w}_{(i),k}$ for $k = j + 1, \ldots, M_{(i)}$.
4. *Segment merge:* Given a segmentation $\mathbf{y}_{(i)} = (M_{(i)}, \mathbf{w}_{(i)})$, propose $M_{(i)}{}^* = M_{(i)} - 1$ by merging the j^{th} segment and $j + 1^{th}$ segment, where j is sampled uniformly from $[1, M - 1]$. Set $\mathbf{w}^*_{(i),k} = \mathbf{w}_{(i),k}$ for $k = 1, \ldots, j - 1$, and $\mathbf{w}^*_{(i),k-1} = \mathbf{w}_{i,k}$ for $k = j + 1, \ldots, M_i$.

Most structural motifs we discuss in this article have regular arrangement of the secondary structure elements so that the state transitions are deterministic or almost deterministic. Therefore, the operator for *state transition* can be removed and *segment split or merge* can be greatly simplified. There might be some cases that the cross-chain interactions are also stochastic in a complex quaternary motif. Then two additional operators are necessary, including *segment join* (adding an interaction edge in the protein structure graph) and *segment separate* (deleting an interaction edge in the graph). The detailed steps are similar to *state transition*, and we omit detailed discussion here.

4.2. Testing phase

Given the test protein sequences without resolved structures, we need to predict the best segmentation that yields the highest conditional likelihood. Similar to the training phase, it is an optimization problem involving search in multiple-dimensional space. Since it is computationally prohibitive to search over all possible solutions using the traditional optimization methods, simulated annealing with reversible jump MCMC is used. It has been shown theoretically and empirically to converges on the global optimum (Andrieu et al., 2000). Algorithm 2 shows the detailed description of reversible jump MCMC simulated annealing. β is a parameter to control the temperature reduction rate and set to 0.5 in our experiments.

Algorithm 2 Description of reversible jump MCMC simulated annealing

```
     Input: initial value of $y_0$, Output: optimized assignment of $\mathbf{y}$
1.   Set $\hat{\mathbf{y}}_0$.
2.   For $t \leftarrow 1$ to $\infty$ do:
       2.1 $T \leftarrow \beta t$. If $T = 0$ return $\hat{\mathbf{y}}$
       2.2 Sample a value from $\mathbf{y}^{new}$ using the reversible jump MCMC algorithm as described in
           Section 4.1. $\nabla E = \Phi(\mathbf{y}^{new}) - \Phi(\hat{\mathbf{y}})$
       2.3 if $\nabla E > 0$, then set $\hat{\mathbf{y}} = \mathbf{y}^{new}$; otherwise set $\hat{\mathbf{y}} = \mathbf{y}^{new}$ with probability $\exp(\nabla E/T)$
3.   Return $\hat{\mathbf{y}}$
```

5. FEATURE EXTRACTION

The conditional graphical models provide an expressive framework to capture the structural properties of target motifs characterized by both local interactions, inter-chain and intra-chain interactions. They enjoy the advantages of the original CRF model so that any type of informative features, either overlapping or long-range correlations, can be used conveniently. Similar to other applications, the choice of feature function f_k plays an essential role in accurately predicting the structural motifs.

From the perspective of graph topology, two types of features can be defined, i.e., *node features*, which capture the properties of an individual structural component, and *pairwise features*, which model the chemical-bonding between pairs of structural components that are close in three-dimensional spaces. One common approach to define the feature function in the CRF-like models is factorization. For example, for all the models we discussed above, we can define the node features $f_{(L^*,S^*)}(\mathbf{x}, s_i, d_{i-1}, d_i)$ as follows:

$$f_{(L^*,S^*)}(\mathbf{x}, s_i, d_{i-1}, d_i) = f_k^{'}(\mathbf{x}, d_{i-1}, d_i)\delta(d_i - d_{i-1}, L^*)\delta(s_i, S^*), \tag{10}$$

where $L^* \in [l_{min}^{S^*}, l_{max}^{S^*}], S^* \in \mathcal{S}$, and $\mathcal{S}$ is the set of state assignments; δ is the indicator function; and $f_k^{'}(\mathbf{x}, d_{i-1}, d_i)$ is the feature defined over the observed subsequences $x_{di-1}+1x_{di-1} \ldots x_{di}$. Similarly, the pairwise features $g_{(L_a^*,S_a^*),(L_b^*,S_b^*)}(\mathbf{x}, s_j^{(i)}, d_{j-1}^{(i)}, d_j^{(i)}, s_q^{(p)}, d_{q-1}^{(p)}, d_q^{(p)})$, can be factorized as follows:

$$\begin{aligned} & g_{(L_a^*,S_a^*),(L_b^*,S_b^*)}(\mathbf{x}, s_j^{(i)}, d_{j-1}^{(i)}, d_j^{(i)}, s_q^{(p)}, d_{q-1}^{(p)}, d_q^{(p)}) = \\ & \quad g'(\mathbf{x}, d_{j-1}^{(i)}, d_j^{(i)}, d_{q-1}^{(p)}, d_q^{(p)})\delta(d_j^{(i)} - d_{j-1}^{(i)}, L_a^*)\delta(d_q^{(p)} - d_{q-1}^{(p)}, L_b^*)\delta(s_j^{(i)}, S_a^*)\delta(s_q^{(p)}, S_b^*), \end{aligned}$$

where $g'(\mathbf{x}, d_{j-1}^{(i)}, d_j^{(i)}, d_{q-1}^{(p)}, d_q^{(p)})$ is the feature defined over a pair of subsequences $x_{d_{j-1}+1}^{(i)} x_{d_{j-1}}^{(i)} \ldots x_{d_j}^{(i)}$ and $x_{d_{q-1}}^{(p)} + 1x_{d_{q-1}}^{(p)} \cdots x_{d_q}^{(p)}$.

The features (f' and g') useful for predicting the structural motifs can be summarized as two types: one is *common features*, which capture the common characteristics of protein structures, such as physi-chemical properties of amino acid or the propensity that the two residues can form hydrogen bonds in the β-sheets; the other is *signal features*, which are unique to the target structural motif but require domain expertise. Our experiments and studies show that the signal features usually provide the most discriminative information about the target motif and are given higher weights in the learned models. However, it is time-consuming to get those signal features: generally it takes years for the biologists to acquire the related domain knowledge. Sometimes, our current understanding of the target motif (e.g., the double-barrel trimer motif) is not enough to summarize any reliable signal patterns, in which case the common features could be a reasonable backup.

The common node features we use to predict the structural motifs in our experiments include: the maximal, minimal and mean of the secondary structure prediction scores for each position in the subsequence, the physicochemical properties, such as Kyte-Doolittle hydrophobicity score, solvent accessibility and ionizable score.[2] The pairwise features we find useful for β-sheet related motifs or folds include the side chain alignment scores based on the different propensities to form a hydrogen bond depending on whether the side-chains are buried or exposed (Bradley et al., 2002), the propensity of the different pairs of amino acids to form parallel or anti-parallel β-sheets (Steward and Thornton, 2002), and the distance between the interacting pairs. The signal features for the target motifs are usually represented via the sequence, that is, the biologists have summarized or hypothesized some sequence templates based on the experiments or observations from the known positive proteins. Therefore, we can use the template matching scores as features. Similarly, we can also build regular expression templates or statistical profiles to capture the sequence conservations that contribute to the stability of the whole structures. Table 1 summarizes the features we use in the experiments.

6. EXPERIMENTS

To evaluate the effectiveness of our approach, we use four structural motifs as examples, including two repetitive motifs: β-helixes, an elongated helix-like structure with a series of progressive stranded coilings

[2] The score tables of these properties can be accessed at *www.cgl.ucsf.edu/chimera/1.2065/docs/UsersGuide/midas/hydrophob.html, http://prowl.Rockefeller.edu/aainfo/access.html*

Table 1. Feature Definition for Segment $w_i = \{s_i, d_i\}$ and $w_j = \{s_j, d_j\}$

	Feature type	Semantics	Examples
Common features	Node features	Maximum predicted 2nd structure scores	$\max_{t \in [d_i, d_{i+1}]} P_{\beta-\text{sheet}}(x_t)$
		Minimum predicted 2nd structure scores	$\min_{t \in [d_i, d_{i+1}-1]} P_{\beta-\text{sheet}}(x_t)$
		Averaged predicted 2nd structure scores	$\sum_{t \in [d_i, d_{i+1}-1]} P_{\beta-\text{sheet}}(x_t) / (d_{i+1} - d_i)$
		Segment length	$d_{i+1} - d_i$
		Averaged physicochemical property scores (hydrophobicity, solvent accessibility, ionizable)	$\sum_{t \in [d_i, d_{i+1}-1]} \text{S}_{\text{ionic}}(x_t) / (d_{i+1} - d_i)$
	Pairwise features	Side-chain alignment scores (buried (B) or exposed (E)) [10]	$\sum_{t \in [0, \ell]} I(x_i = \text{buried}) S_B(x_{t+d_i}, x_{t+d_j}) + I(x_i = \text{exposed}) S_E(x_{t+d_i}, x_t + d_j)$
		Parallel/anti-parallel β-sheet alignment score [38]	$\sum_{t \in [0, \ell]} S_{\text{parallel}}(x_{t+d_i}, x_{t+d_j})$
Signal features	Triple-β spirals	REM for B1-strand side-chain alternating patterns	$x_{d_i} \ldots x_{d_{i+1}} = \sim \text{X}\Upsilon\Phi\text{X}\Psi\text{XX}$
		REM for B2-strand side-chain alternating patterns	$x_{d_i} \ldots x_{d_{i+1}-1} = \sim \text{XX}\Phi\text{X}\Phi\text{X}\Psi\text{X}$
		B1 (B2) alignment profile matching	$P_{\text{HMMER - B1}}(x_{d_i} \ldots x_{d_{i+1}-1})$
	Double-barrel trimer	Max β-turn score (6 type: I, II, VIII, I′, II′,VIa, VIb, and IV) [18]	$\max_{t \in [d_i, d_{i+1}-1]} S_{\text{type I } \beta-\text{turn}}(x_t)$
	β-helix	REM for B2-T2-B3 side-chain alternating pattern	$x_{d_i} \ldots x_{d_{i+1}-1} = \sim \Phi\text{X}\Phi\text{XX}\Psi\text{X}\Phi\text{X}$
		B1 (B2-T2-B3) profile HMM alignment matching	$P_{\text{HMMER - B1}}(x_{d_i} \ldots x_{d_{i+1}-1})$
	Leucine-rich repeats	REM for LLR side-chain alternating pattern	$x_{d_i} \ldots x_{d_{i+1}-1} = \sim \text{XXXLXXLX[LV]XXXXX}$
		B1 (B2-T2-B3) profile LLR alignment matching	$P_{\text{HMMER - LLR}}(x_{d_i} \ldots x_{d_{i+1}-1})$

Notation: $\ell = d_{i+1} - d_i$, $\Upsilon \in$ {P, G, A, F, S, L}, $\Phi \in$ {L, I, M, V, T, S, F, A}, $\Phi \notin$ {C, E, H, P, Q, W}, X match any amino acid. "$= \sim$" indicates that the string matches the regular expression. REM, regular expression matching.

that are responsible for binding the O-antigens, and leucine-rich repeats (LLR), a solenoid-like regular arrangement of β-strand and α-helix that involve in various protein-protein interaction processes (Kobe and Deisenhofer, 1994); and two quaternary motifs: TBS, a virus fiber that initiate the binding to the cellular receptor molecule, and the double-barrel trimer (DBT), which comprises the virus capsids to protect the DNA or RNA. We choose them specifically because they are good examples of structural motifs residing in the twilight zone of sequence similarity, perform important functions, and more importantly, all these motifs find common existence in viruses infecting different species. On one hand, identifying more examples of those motifs might reveal important evolution relationships among the viral proteins; on the other hand, it is extremely difficult to experimentally determine the membership proteins due to their subcellular locations and the complexity of the structures. Therefore the unavailability of sufficient training data and low sequence conservation pose great challenges to computationally predict those motifs.

Our goal is to identify potential proteins taking the target motif from the whole collection of protein sequences without resolved structures. It can be treated a ranking problem and so the evaluation measure is to see whether we can rank the held-out known membership proteins higher than the negative examples in cross-validation. To construct negative examples in the training set, we follow the standard approaches in bioinformatics, i.e., building the PDB-minus dataset, which consists of all protein sequences with known structures in PDB (July 2006 version) (Berman et al., 2000) with less than 25% similarity to each other and no less than 40 residues in length, resulting in 2810 chains with 430,927 residues. Since we search for proteins sharing similar structures without sequence similarity, a leave-family-out cross-validation was performed to avoid overfitting. For each cross, positive proteins from the same protein family are placed in the test set while the remainder are placed in the training set. Similarly, the PDB-minus set was also randomly partitioned into subsets, one of which is placed in the test set while the rest serve as the negative training examples. Since negative data dominate the training set, we select a subset of negative sequences (about five times the size of the positives) that are most similar to the positive examples in sequence identity so that the models can learn a better decision boundary than randomly sampling. When inferencing, we stop the iterative searching algorithm when the differences of loglikelihood is less than 0.001 or the iteration time is larger than 5000; the number of sampling steps T in the contrastive divergence is set to 5; the number of iterations in simulated annealing is 500. The score is the log ratio between the probability of the best segmentation and that of the whole sequence as one segment in null state. To determine whether a protein sequence has a particular fold, we define the score ρ as the normalized log ratio of the probability for the best segmentation to the probability of the whole sequence in a null state (non-β-helix or non-LLR). We compare our results with Threader, a threading algorithm which minimizes the potential function based on physical forces, and HMMER, a general motif detection algorithm using profile HMMs (Karplus et al., 1998). The input to HMMER can be the structural alignments using CE-MC (Guda et al., 2004) or purely sequence-based alignments by CLUSTALW (Thompson et al., 1994). We also compare our results on β-helix with BetaWrap, the state-of-art algorithm designed specifically to predict the β-helices.

6.1. β-Helices

There currently exist 14 protein sequences with three-stranded right-hand β-helix whose crystal structures have been deposited in PDB. The sequence similarity between those 14 proteins is less than 25%, which falls in the "twilight" zone where most current algorithms fail. 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). Table 2 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 conditional graphical models can successfully score all known β-helices higher than non β-helices in PDB, significantly better than Threader, HMMER and BetaWrap, the stat-of-art method for predicting the β-helices fold. Our algorithm also demonstrates success in locating each repeat in the known β-helix proteins. In Table 3, we cluster the proteins into three different groups according to the segmentation results and show examples of the predicted segmentation in each group. We also hypothesize potential β-helix proteins from the whole sequence database, i.e., Swiss-Prot, using conditional graphical models. The full list can be accessed at *www.cs.cmu.edu/~yanliu/SCRF.html*. Up to now, three proteins in our predicted list have been resolved of structures and confirmed having the β-helix motif.

Table 2. Scores and Rank for the Known Right-Handed β-Helices by HMMER Using Structural Alignment, HMMER Using Sequence Alignment, Threader, BetaWrap, and the Conditional Graphical Models (CGMs)

		Struct-based	*HMMs*	*Seq-based*	*HMMs*	*Threader,*	*BetaWrap*		*CGM*	
SCOP family	*PDB-id*	*Bit score*	*Rank*	*Bit score*	*Rank*	*rank*	*Score*	*Rank*	*ρ-score*	*Rank*
P.69 pertactin	1DAB	−73.6	3	−163.4	75	24	−17.84	1	10.17	1
Chondroitinase B	1DBG	−64.6	5	−171.0	55	47	−19.55	1	13.15	1
Glutamate synthase	1EAO	−85.7	65	−109.1	72	N/A	−24.87	N/A	6.21	1
Pectin methylesterase	1QJV	−72.8	11	−123.3	146	266	−20.74	1	6.12	1
P22 tailspike	1TYU	−78.8	30	−154.7	15	2	−20.46	1	6.71	1
Iota-carrageenase	1KTW	−81.9	17	−173.3	121	10	−23.4	N/A	8.07	1
Pectate lyase	1AIR	−37.1	2	−133.6	35	45	−16.02	1	16.64	1
	1BN8	180.3	1	−133.7	37	76	−18.42	3	13.28	2
	1EE6	−170.8	852	−219.4	880	228	−16.44	2	10.84	3
Pectin lyase	1IDJ	−78.1	14	−178.1	257	6	−17.99	2	15.01	2
	1QCX	−83.5	28	−181.2	263	6	−17.09	1	16.43	1
Galacturonase	1BHE	−91.5	18	−183.4	108	18	−18.80	1	20.11	3
	1CZF	−98.4	43	−188.1	130	5	−19.32	2	40.37	1
	1RMG	−78.3	3	−212.2	270	27	−20.12	3	23.93	2

Notice that the bit scores from HMMER are not directly comparable.

6.2. *Leucine-rich repeats*

There are 41 LLR proteins with known structure in PDB, covering two super-families and 11 families in SCOP. The LLR motif is relatively easy to detect due to its sequence conservations with many leucines and short insertions. Therefore it would be more interesting to discover new LLR proteins with much less sequence identity to previous known proteins. We select one protein in each family as representative and see if our model can identify LLR proteins across families. Table 4 lists the output scores by different methods and the rank for the LLR proteins. In general, LLR is easier to identify than the β-helices. Again, the conditional graphical models perform much better than other methods. In addition, the predicted segmentation by our model is close to prefect match for most LLR proteins (some examples are shown in Fig. 4).

6.3. *Triple β-spirals*

Up to now there are only three crystallized structures with the TBS motif deposited in PDB. The sequence similarity between these three proteins are lower than 20%. Given the limited number of training examples and low sequence conservation, we can see that it is extremely challenging to predict this motif. Table 5 shows the scores and rank of different algorithms for the known TBS proteins. The conditional graphical models perform much better than other algorithms for this difficult task. Figure 5 shows the histogram of scores and segmentation results predicted by our conditional graphical models for the TBS proteins and the non-TBS proteins. We can observe a relatively clear boundary that separates the TBS proteins from the rest. Of all the proteins that were scored higher than 0 in the PDB-minus set, there are 58 proteins from α class, 45 from β class, 51 from α/β class, 72 from $\alpha + \beta$ class, four from α and β class, and six from membrane class. We also hypothesize potential TBS proteins from the Swiss-Prot database using our model. The whole list can be accessed at *www.cs.cmu.edu/~yanliu/swissprot_list.xls*.

6.4. *Double-barrel trimer*

The DBT is a protein fold that has been found in the coat proteins from several kinds of viruses. It consist of two eight-stranded jelly rolls, or β-barrels (Benson et al., 2004). The layout of the eight β-strands is known, but the specific chemical bonding that maintains the stability of the motif are unclear due to the low resolution of crystallization structures. From Table 6, we can see that it is extremely difficult to predict the DBT motif. Our method is able to give higher ranks for three of the four known DBT proteins, although we are unable to reach a clear separation between the DBT proteins and the rest. The results are within our

Table 3. Segmentation Results (Marked by Colors) for the Known Right-Handed β-Helix by the Conditional Graphical Models

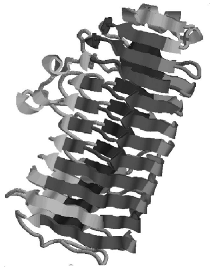
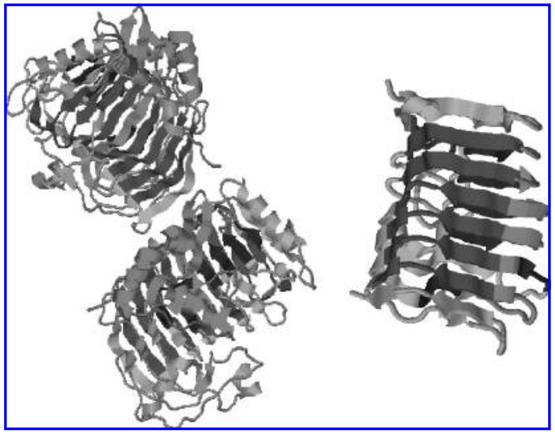
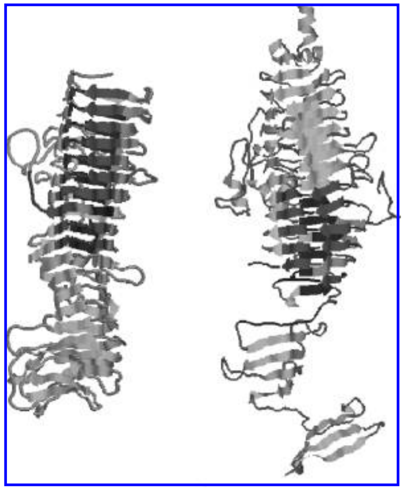
Group	*Perfect match*	*Good match*	*OK match*
No. of missing repeats	0	1–2	3 or more
PDB-ID	1czf	1air, 1bhe, 1bn8, 1dbg, **1ee6**(right), 1idj, **1ktw**(left), 1qcx, 1qjv, 1rmg	**1dab**(left), 1ea0, **1tyu**(right)
			

Table 4. Scores and Rank for the Known Right-Handed Leucine-Rich Repeats (LLR) by HMMER Using Sequence Alignment, HMMER Using Structural Alignment, Threader, and Conditional Graphical Models (CGM)

		Seq-based	*HMMs*	*Struct-based*	*HMMs*	*Threader*	*CGM*	
SCOP Family	*PDB-ID*	*Bit score*	*Rank*	*Bit Score*	*Rank*	*Rank*	*ρ-score*	*Rank*
28-residue LRR	1A4Y	−125.5	4	−76.7	1	457	127.8	1
Rna1p (RanGAP1)	1YRG	−95.4	1	−81.1	1	181	64.3	1
Cyclin ACDK2-associated p19	1FQV	−163.3	89	−111.4	10	398	77.1	1
Internalin LRR domain	1O6V	−62.8	1	−0.7	1	306	116.5	1
Leucine rich effector	1JL5	−86.7	1	−26.5	1	46	187.5	1
Ngr ectodomain-like	1P9A	−120.0	9	−68.6	1	16	105.0	1
Polygalacturonase inhibiting protein	1OGQ	−155.0	32	−18.2	1	284	66.4	1
Rab geranylgeranyltransferase alpha-subunit	1DCE	−145.4	16	−59.7	1	35	17.4	1
mRNA export factor	1KOH	−153.9	42	−91.7	1	177	37.1	1
U2A′-like	1A9N	−280.9	861	−151.4	478	62	55.1	1
L domain	1IGR	−150.0	46	−107.1	249	67	8.2	1

For CGM, ρ-score = 0 for all non-LLR proteins.

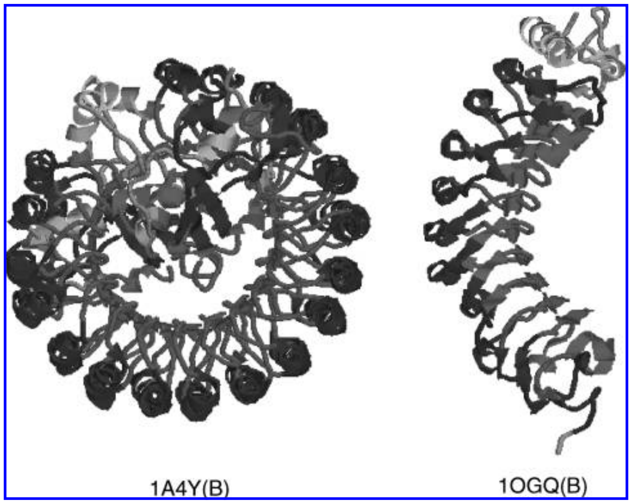
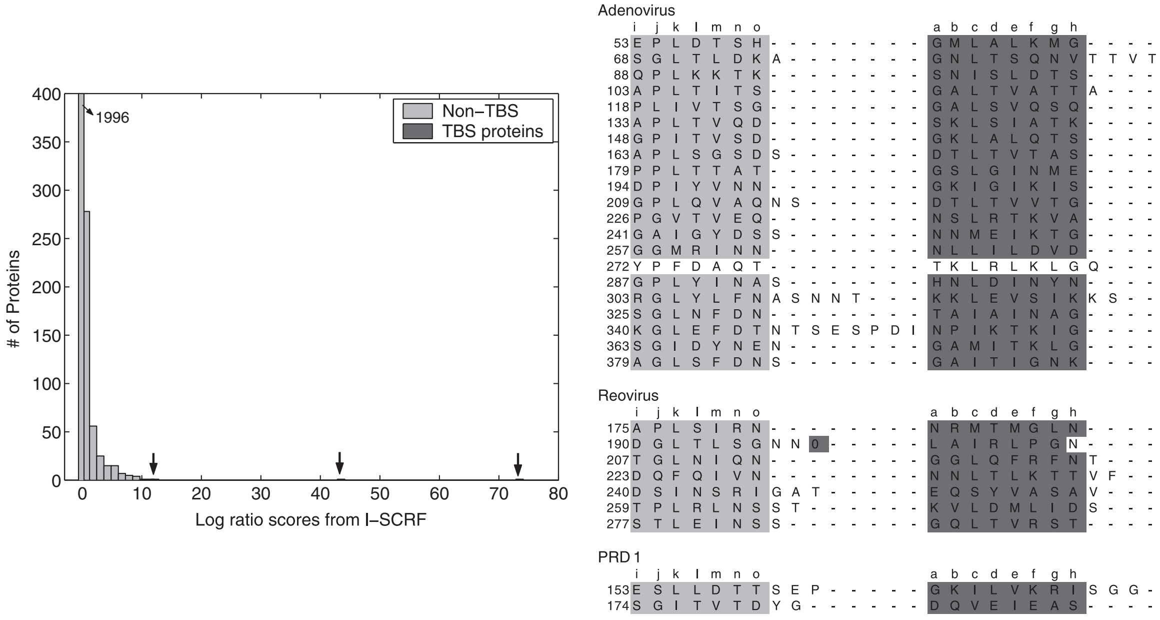


FIG. 4. Segmentation results for example LLR proteins by the conditional graphical models (light grey denotes the fix-template motif and dark grey denotes insertions).

Table 5. Scores and Rank for the Known triple β-Spirals by HMMER Using Sequence Alignment, HMMER Using Structural Alignment, Threader, and Conditional Graphical Models (CGM)

		Seq-based	*HMM*	*Struct-based*	*HMM*	*Threader*	*CGM*	
SCOP family	*PDB-ID*	*Score*	*Rank*	*Score*	*Rank*	*Rank*	*Score*	*Rank*
Adenovirus	1QIU	−343.9	11	−225.5	7	26	74.1	1
Reovirus	1KKE	7.9	1	−294.3	2	242	11.6	1
PRD1	1YQ8	−6.7	7	−399.4	194	928	43.4	1

Notice that the scores from the HMMER are not directly comparable on different proteins.



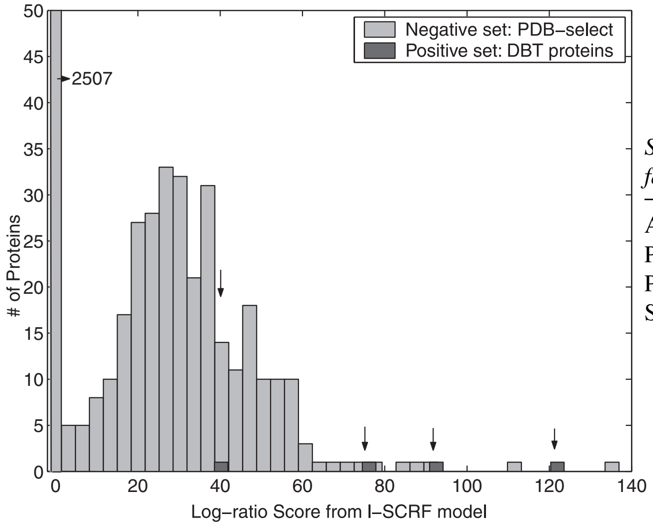
```
Adenovirus
    i j k l m n o                 a b c d e f g h
 53 E P L D T S H - - - - - - - - G M L A L K M G - - - -
 68 S G L T L D K A - - - - - - - G N L T S Q N V T T V T
 88 Q P L K K T K - - - - - - - - S N I S L D T S - - - -
103 A P L T I T S - - - - - - - - G A L T V A T T A - - -
118 P L I V T S G - - - - - - - - G A L S V Q S Q - - - -
133 A P L T V Q D - - - - - - - - S K L S I A T K - - - -
148 G P I T V S D - - - - - - - - G K L A L Q T S - - - -
163 A P L S G S D S - - - - - - - D T L T V T A S - - - -
179 P P L T T A T - - - - - - - - G S L G I N M E - - - -
194 D P I Y V N N - - - - - - - - G K I G I K I S - - - -
209 G P L Q V A Q N S - - - - - - D T L T V V T G - - - -
226 P G V T V E Q - - - - - - - - N S L R T K V A - - - -
241 G A I G Y D S S - - - - - - - N N M E I K T G - - - -
257 G G M R I N N - - - - - - - - N L L I L D V D - - - -
272 Y P F D A Q T - - - - - - - - T K L R L K L G Q - - -
287 G P L Y I N A S - - - - - - - H N L D I N Y N - - - -
303 R G L Y L F N A S N N T - - - K K L E V S I K K S - -
325 S G L N F D N - - - - - - - - T A I A I N A G - - - -
340 K G L E F D T N T S E S P D I N P I K T K I G - - - -
363 S G I D Y N E N - - - - - - - G A M I T K L G - - - -
379 A G L S F D N S - - - - - - - G A I T I G N K - - - -

Reovirus
    i j k l m n o                 a b c d e f g h
175 A P L S I R N - - - - - - - - N R M T M G L N - - - -
190 D G L T L S G N N 0 - - - - - L A I R L P G N - - - -
207 T G L N I Q N - - - - - - - - G G L Q F R F N T - - -
223 D Q F Q I V N - - - - - - - - N N L T L K T T V F - -
240 D S I N S R I G A T - - - - - E Q S Y V A S A V - - -
259 T P L R L N S S T - - - - - - K V L D M L I D S - - -
277 S T L E I N S S - - - - - - - G Q L T V R S T - - - -

PRD 1
    i j k l m n o                 a b c d e f g h
153 E S L L D T T S E P - - - - - G K I L V K R I S G G -
174 S G I T V T D Y G - - - - - - D Q V E I E A S - - - -
```

FIG. 5. (**Left**) Histograms of the cross-validation scores generated by the conditional graphical models on positive examples, i.e., known triple β-spirals (dark bar with arrow indicator) and negative examples, i.e., PDB-select set (light bars). All three held-out proteins score higher than non-TBS proteins. (**Right**) Segmentation results by the conditional graphical models for the known TBS proteins. Predicted B1 strands are shown in light bar and predicted B2 strands in dark bar.

TABLE 6. (LEFT) HISTOGRAMS OF THE CROSS-VALIDATION SCORES GENERATED BY CONDITIONAL GRAPHICAL MODELS ON POSITIVE EXAMPLES, I.E., KNOWN DOUBLE-BARREL TRIMERS (RED BAR WITH ARROW INDICATOR) AND NEGATIVE EXAMPLES, I.E., PDB-SELECT SET (GREEN BARS). (RIGHT) RANKING FOR THE KNOWN DOUBLE-BARREL TRIMER BY HMMER USING SEQUENCE ALIGNMENT, HMMER USING STRUCTURAL ALIGNMENT, THREADER, AND CONDITIONAL GRAPHICAL MODELS (CGMS)



SCOP family	*Seq-based HMMs*	*Struct-based HMMs*	*Threader*	*CGMs*
Adenovirus	12	14	>385	87
PRD1	84	107	323	**8**
PBCV	92	8	321	**3**
STIV	218	70	93	**2**

expectation because the lack of signal features and poor understanding about the inter-chain interactions make the prediction significantly harder. Of all the proteins scored higher than 0 in the PDB-minus set, there are 45 proteins from α class, 37 from β class, 88 from α/β class, 28 from $\alpha+\beta$ class, 14 from α and β class, and seven from membrane class. We believe more improvement can be achieved by combining the results from multiple algorithms.

7. DISCUSSION

In this paper, we discuss a graphical model approach for protein structural motif prediction. This task can be seen as an extension of previous studies in protein structure prediction, but differs in two aspects: First, our task comes directly from the needs of biologists in their experiments or studies. The structural motifs that we examined in the paper have been studied for decades by our collaborators. The number of membership proteins with resolved structures are very limited, although these motifs are believed to exist commonly in nature. By identifying more examples of the target motif in genome-wide sequence databases, we can help the biologists to reduce their searching space and thus speed the verification of their hypothesis. Second, our task is much more difficult than the common fold classification, as discussed in Cheng and Baldi (2006) and Ding and Dubchak (2000), because we are focusing on less representative structural motifs, i.e., those with much fewer positive examples and less sequence conservation. In other words, the patterns that we are trying to identify have not been reflected clearly in the training sequences, which motivates us to develop a sophisticated model so that domain knowledge can be easily incorporated, rather than applying a simple classifier. In other words, our models can be used in the classical motif identification or fold recognition task, but its advantages are demonstrated best when predicting those difficult structural motifs.

8. CONCLUSION

In this paper, we present a new and effective framework, the conditional graphical models, for predicting the protein structural motifs. We demonstrate that, by examining the structural properties of the target motifs and incorporating them into our models, we are able to solve the problem effectively. Compared with previous models for structured prediction, our models are more general in that they are able to capture the long-range interactions, either between segments within one chain or on different chains. In addition, we develop the efficient learning and inference algorithms using the reversible jump MCMC sampling. For future work, it would be interesting to combine the conditional graphical models with active learning, in which we can automatically bootstrap negative features from the motif databases using false positive examples.

ACKNOWLEDGMENTS

We thank Jonathan King for his input and biological insights, and John Lafferty and Eric P. Xing for their help on deriving the models. This material is based upon work supported by the National Science Foundation (grant 0225656).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aloy, P., and Russell, R. 2003. Interprets: protein interaction prediction through tertiary structure. *Bioinformatics* 19, 161–162.

Altschul, S., Madden, T., Schaffer, A., et al. 1997. Gapped BLAST and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Andrieu, C., de Freitas, N., and Doucet., A. 2000. Reversible jump MCMC simulated annealing for neural networks. *Proc. UAI-00,* 11–18.

Bateman, A., Coin, L., Durbin, R., et al. 2004. The PFAM protein families database. *Nucleic Acids Res.* 32, 138–141.

Benson, S., Bamford, J., Bamford, D., et al. 2004. Does common architecture reveal a viral lineage spanning all three domains of life? *Mol Cell.* 16, 673–685.

Berman, H., Westbrook, J., Feng, Z., et al. 2000. The protein data bank. *Nucleic Acids Res.*, 28, 235–242.

Bourne, P.E., and Weissig. H. 2003. *Structural Bioinformatics: Methods of Biochemical Analysis*. Wiley-Liss, New York.

Boys, R.J., and Henderson. D.A. 2001. A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Comp. Sci. Statist.* 33, 35–49.

Bradley, P., Cowen, L., Menke, M., et al. 2001. Predicting the beta-helix fold from protein sequence data. *Proc. RECOMB'01*.

Bradley, P., Kim, P. S., and Berger. B. 2002. Trilogy: discovery of sequence-structure patterns across diverse proteins. *Proc. RECOMB'02*.

Cheng, J., and Baldi. P. 2006. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 22, 1456–1463.

Ding, C.H., and Dubchak. I., 2000. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358.

Do, C., Gross, S., and Batzoglou. S. 2006. Contralign: discriminative training for protein sequence alignment. *RECOMB'06*.

Durbin, R., Eddy, S., Krogh, A., et al. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, New York.

Fischer, D. 2000. Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*

Fleishman, S.J., and Ben-Tal, N. 2006. Progress in structure prediction of a-helical membrane proteins. *Curr. Opin. Struct. Biol.* 16, 496–504.

Fuchs, P., and Alix,. A, 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins* 59, 828–839.

Green. P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.

Guda, C., Lu, S., Sheeff, E.D., Bourne, P.E., Shindyalov, I.N. 2004. CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res* (in press).

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

Huelsenbeck, J., Larget, B., and Alfaro, M. 2004. Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. *Mol. Biol. Evol.* 6, 1123–1133.

Jones, D., Taylor, W., and Thornton, J. 1992. A new approach to protein fold recognition. *Nature* 358, 86–89.

Kamisetty, E.X.H., and Langmead, C. 2007. Free energy estimates of all-atom protein structures using generalized belief propagation. *RECOMB'07*.

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

Kelley, L., MacCallum, R., and Sternberg, M. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 229, 499–520.

Kobe, B., and Deisenhofer, J. 1994. The leucine-rich repeat: a versatile binding motif. *Trends Biochem. Sci.* 10, 415–421.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. ICML'01*.

Lu, L., Lu, H., and Skolnick, J. 2002. Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins.* 49, 350–364.

Menke, M., Scanlon, E., King, J. et al. 2004. Wrap-and-pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils. *Proc. RECOMB'04*.

Murray, I., and Ghahramani, Z. 2004. Bayesian learning in undirected graphical models: approximate mcmc algorithms. *Proc. UAI-04* 392–399.

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

Orengo, C., Michie, A., Jones, S., et al. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108.

Rooman, M., and Wodak, S. 1995. Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* 8, 849–858.

Sander, O., Sommer, I., and Lengauer, T. 2006. Local protein structure prediction using discriminative models. *BMC Bioinform.* 7, 14.

Sarawagi, S., and Cohen, W.W. 2004. Semi-Markov conditional random fields for information extraction. *Proc. NIPS'2004*.

Scanlon, E.L., 2004. Predicting the triple beta-spiral fold from primary sequence data [M.S. Thesis]. Massachusetts Institute of Technology, Cambridge, MA.

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

Taskar, B., Guestrin, C., and Koller., D. 2003. Max-margin Markov networks. *Proc. NIPS'03*.

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. *Nucleic Acids Res.* 22, 4673–4680. 1994.

van Raaij, M., Mitraki, A., Lavigne, G., et al. 1999. A triple beta-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein. *Nature* 401, 935–938.

Venclovas, C., Zemla, A., Fidelis, K., et al. 2003. Assessment of progress over the CASP experiments. *Proteins* 53, 585–595.

von Ohsen, N., Sommer, I., and Zimmer, R. 2003. Profile-profile alignment: a powerful tool for protein structure prediction. *Pac. Symp. Biocomput.*

Weigele, P.R., Scanlon, E., and King. J. 2003. Homotrimeric, β-stranded viral adhesins and tail proteins. *J. Bacteriol.* 185, 4022–4030.

Welling, M., and Hinton, G.E., 2002. A new learning algorithm for mean field Boltzmann machines. *Proc. ICANN '02* 351–357.

Westhead, D., Slidel, T., Flores, T. et al. 1999. Protein structural topology: automated analysis and diagrammatic representation. *Protein Sci.* 8, 897–904.

Yang, J., Liu, Y., Xing, E. P., et al. 2007. Harmonium-based models for semantic video representation and classification. *Proc. SIAM Int. Conf. Data Mining*.

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

Address reprint requests to:
Dr. Yan Liu
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

E-mail: liuya@us.ibm.com

This article has been cited by:

1. R. Day, K. P. Lennox, D. B. Dahl, M. Vannucci, J. W. Tsai. 2010. Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. *Bioinformatics* **26**:24, 3059-3066. [CrossRef]
2. A. Kumar, L. Cowen. 2010. Recognition of beta-structural motifs using hidden Markov models trained with simulated evolution. *Bioinformatics* **26**:12, i287-i293. [CrossRef]
3. M. Menke, B. Berger, L. Cowen. 2010. Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system. *Proceedings of the National Academy of Sciences* **107**:9, 4069-4074. [CrossRef]