Predicting Protein Folds with Structural Repeats Using a Chain Graph Model

Yan Liu[†] YANLIU@CS.CMU.EDU
Eric P. Xing[†♯] EPXING@CS.CMU.EDU
Jaime Carbonell[†] JGC@CS.CMU.EDU
LTI[†] and CALD[♯], School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

Protein fold recognition is a key step towards inferring the tertiary structures from amino-acid sequences. Complex folds such as those consisting of interacting structural repeats are prevalent in proteins involved in a wide spectrum of biological functions. However, extant approaches often perform inadequately due to their inability to capture long-range interactions between structural units and to handle low sequence similarities across proteins (under 25% identity). In this paper, we propose a chain graph model built on a causally connected series of segmentation conditional random fields (SCRFs) to address these issues. Specifically, the SCRF model captures long-range interactions within recurring structural units and the Bayesian network backbone decomposes cross-repeat interactions into locally computable modules consisting of repeat-specific SCRFs and a model for sequence motifs. We applied this model to predict β-helices and leucine-rich repeats, and found it significantly outperforms extant methods in predictive accuracy and/or computational efficiency.

1. Introduction

The tertiary structures of proteins play key roles in determining the function, activity, stability and subcellular localization of proteins, and the mechanisms of protein-protein interactions in cells. An important issue in inferring tertiary structures from amino-acid sequences is how to accurately identify protein folds arising from typical spatial arrangements of well-defined secondary structures that can be recognized from the sequence. Given the putative protein folds present in a protein, the backbone of the tertiary structure can be more easily inferred. More importantly, these folds may also serve as key indicators for certain functional sites. *In silico* protein fold recognition seeks to predict whether a given protein sequence contains a putative structural fold (usually represented by a training set of instances of this fold) and if so, locate its exact position within the sequence.

To date, there has been significant progress in predicting certain types of simple *well-defined* supersecondary structures, such as $\alpha\alpha$- and $\beta\beta$-hairpins, based on their primary sequences using rule-based algorithms or hidden Markov models (Durbin et al., 1998). However, predicting *more complex and irregular* protein folds such as those containing highly stochastic (in terms of sequence composition, spacing and ordering) internal structures remains an open problem.

In this paper, we address a special class of the aforementioned complex protein folds—those with repetitive structural motif components, such as the β-helices (Yoder et al., 1993) or the leucine rich repeats (LLR) (Kobe & Deisenhofer, 1994) (Fig.1). These folds are believed to be prevalent in proteins and can 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 & Deisenhofer, 1994). Identifying these folds remains a challenge because of the presence of many complex and irregular features in their structure—for example, long-range interactions between their build-blocks (i.e., structural motifs) separated by an unknown number of spacers (i.e., amino acid insertions), low sequence similarities (less than 25%) between recurring motifs within the same protein and across multiple proteins, and non-conserved insertions of variable lengths across different proteins.

The traditional approaches for protein fold prediction search the database using PSI-BLAST (Altschul et al., 1997) or match against an HMM profile built from

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

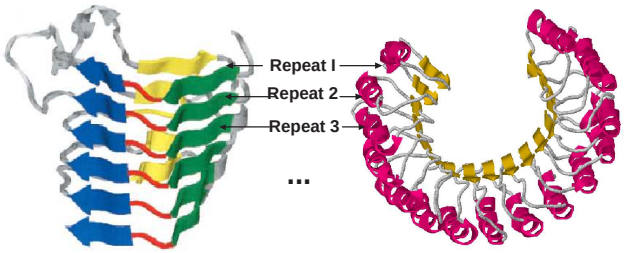

Figure 1. Typical 3-D structure of proteins with β-helices (left) and leucine-rich repeats (right). In β-helices, there are three strands: B1 (green), B2 (blue) and B3 (yellow) and the conserved T2 turn (red). In LLR, there is one strand (yellow) and insertions with helices (red).

sequences with the same fold (Durbin et al., 1998). These methods work well for simple folds with strong sequence similarities, but fail when the sequence similarity across proteins is poor and/or there exist long-range interactions between elements in the folds. Several more expressive probabilistic models that explicitly capture these structural features have been proposed. Delcher *et al.* introduced probabilistic causal networks for protein secondary structure modeling (Delcher et al., 1993). Recently, Lafferty *et al.* applied kernel conditional random fields (kCRFs) for protein secondary structure prediction (Lafferty et al., 2004); Chu *et al.* extended segmental semi-Markov model under the Baysian framework to predict secondary structures (Chu et al., 2004).

While the aforementioned models have led to some improvements in protein structure prediction, they remain inadequate for complex protein folds containing stochastic arrangement of repeating patterns of motifs and insertions. In these proteins, some motifs are quite conserved in sequences or prefer specific lengths; others might be spatially close enough in 3-D to form hydrogen-bonds, such as two β-strands in a parallel β-sheet and helix pairs in coupled helical motifs. Therefore it is necessary to construct a model that explicitly captures these properties. In this paper, we propose a *chain graph model* based on a "protein structural graph". In this graph, nodes are introduced to represent motifs, insertions or relevant structural states. The edges indicate the interactions between these elements in 3-D. Our chain graph model uses segmentation CRFs (SCRFs) as building blocks to capture the long-range interactions between structural repeats, and also employs a mixture profile model to explore the similarities of recurring motifs within the same protein and across multiple proteins. A Bayesian network backbone decomposes cross-repeat interactions into locally computable modules consisting of repeat-specific SCRFs and the model for sequence motifs. As a result, our model not only can capture rich structure features of complex folds, but is also much more efficient than the previously proposed graphical model for protein fold recognition (Liu et al., 2005). Notice that our model can be understood as an approach for simultaneously classifying and segmenting the protein sequences, whereas most previous work perform classification without examining the fine details of structural arrangement (Ding & Dubchak, 2001).

The rest of the paper is organized as follows, we first define the notation and initial settings for the fold-prediction model. Then we overview the SCRF model which serves as the key building block for our new model. In section 3 we describe a novel chain graph model built upon SCRFs and a sequence motif sub-model. In section 4 we report experimental results on two types of protein folds. We conclude with a brief summary and an outline of future work.

2. Segmentation CRFs for protein fold recognition

2.1. Terminology and notation

Protein folds with structural repeats are defined as repetitive secondary or supersecondary structural units, such as α-helices, β-strands, β-sheets (colored regions is Fig.1), connected by *insertions* of variable lengths, which are mostly short loops and sometimes α-helices or/and β-sheets (gray regions in Fig.1).

A graphical model (GM) can be used to define the probability distribution over all possible structural configurations underlying a given protein sequence. We refer to such a GM as a "protein structural graph" (PSG). Specifically, a PSG is an annotated graph $G = \{V, E\}$, where V is the set of nodes corresponding to the specificities of structural units, such as motifs, insertions or the regions outside the fold (which are unobserved and must be inferred), and the amino acid residues at each position (which are observed and should be conditioned on). E represents the set of edges denoting dependencies between the objects represented by the nodes, such as locational constraints and/or state transitions between adjacent nodes in the primary sequence, or long-range interactions between non-neighboring motifs and/or insertions (see Fig.2 (A)). Note that the latter type of dependencies is unique to our PSG, and is the main cause of its computational complexity. A probabilistic distribution on a graph can be postulated by using the potential functions defined on the *cliques* of nodes induced by the edges in the graph (Hammersley & Clifford, 1971).

Given a protein sequence $\mathbf{x} = x_1 x_2 \dots x_n$, where $x_i \in \{\text{amino acids}\}$ and n is the length of the sequence, a "conditional" PSG is defined as follows. Let $\mathbf{S} = (S_1, S_2, \dots, S_M)$, where $S_i \in \{1, \dots, n\}$ denotes the *ending position* of the i^{th} structural segment. Let $\mathbf{T} = (T_1, T_2, \dots, T_M)$, where $T_i \in \mathcal{T}$ denotes the *label* of the segment and $\mathcal{T}$ is a finite set of structural labels.

Finally, let $M \in \{1, \ldots, m_{\max}\}$ denote the number of possible segments in the protein, where $m_{\max}$ can be specified by domain experts or postulated from the training instances. Under this setup, a value assignment to the nodes $W = \{M, \mathbf{S}, \mathbf{T}\}$ in a PSG defines a unique segmentation and annotation of protein $\mathbf{x}$. With a slight abuse of the notation, we use W_i to represent a segment-specific clique (i.e., $W_i = \{S_{i-1}, S_i, T_i\}$, see Fig.2 (A)) that completely determines the configuration of the i^{th} segment. Likewise, an arbitrary clique $c \in \mathcal{C}_G$ can be represented by W_c. Now, for a given PSG G, the conditional probability of W given the observation $\mathbf{x}$ can be defined as

$$P(W|\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}_G} \Phi(\mathbf{x}, W_c), \qquad (1)$$

where $\mathcal{C}_G$ represents the set of all cliques in G, $\Phi(\cdot)$ is the potential function defined on a clique, and Z denotes the normalization constant. Given a query protein, our goal is to seek the segmentation (i.e. W^{opt}) that optimizes this conditional probability.

2.2. Segmentation conditional random fields

Recently a segmentation CRFs model was proposed for general protein fold recognition (Liu et al., 2005). Following (Lafferty et al., 2001), SCRFs assume that the potential function of interest admits an exponential representation, i.e. $\Phi(\mathbf{x}, W_c) = \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, W_c))$, where $f_k(\cdot)$ denotes a feature defined on cliques c, such as the secondary structure assignments or the length of the segment. Since the spatial topology of regular protein folds is often known *a priori*, a deterministic dependency between states T_i and T_{i+1} results. This leads to a simplification that only the cliques involve in the known long-range interactions need to be considered (e.g., "red" arc in Fig.2 (A)). Therefore we have:

$$P(W|\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^{M} \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}, W_i, W_{\pi_i})), \qquad (2)$$

where W_{π_i} denotes the spatial predecessor (i.e., with small position index) of W_i connected by a "long-range interaction arc". The model parameters λ can be estimated by maximizing the regularized log-loss of the training data using iterative searching algorithms, such as gradient descent or L-BFGS (Minka, 2001). The convexity property guarantees that the root corresponds to the optimal solution.

After the simplification, if the graph G can be viewed as a set of chains, a forward-backward algorithm analogous to the one for the original CRFs (Lafferty et al., 2001) can be applied to compute optimal segmentation and labeling under SCRFs (Liu et al., 2005). In general, the computational cost of SCRFs for the forward-backward probabilities and the Viterbi algorithm is

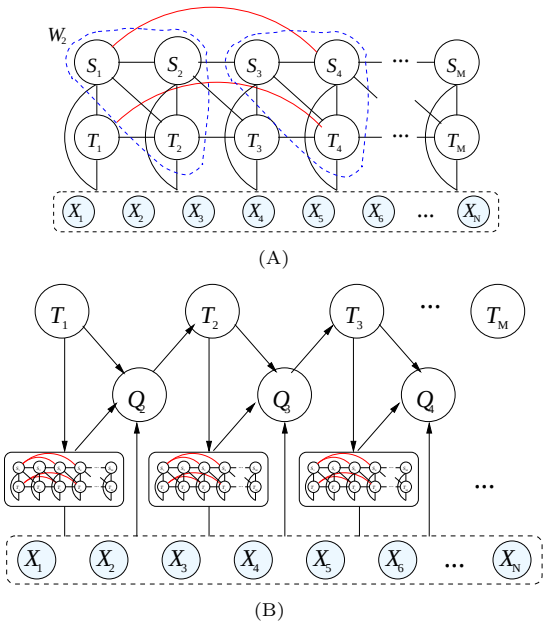

Figure 2. The graphical model representation of protein fold models. A) The SCRF model. Circles represent the state variables, edges represent couplings between the corresponding variables (in particular, long-range interaction between units are depicted by red arcs). The dashed triangles are examples of "segment-specific cliques". The dashed box over x's denote the sets of observed sequence variables. An edge from a box to a node is a simplification of dependencies between the non-boxed node to all the nodes in the box (and therefore result in a clique containing all x's). B) The chain graph model. The directed edges denote conditional dependencies of the child node on the parental nodes. Note that each of the round-cornered boxes represents a repeat-specific component as SCRFs. An edge from the box denote dependencies on the joint configuration of all nodes within the box.

$O(n^3)$. If the possible length of each segment is much smaller than n or fixed, which are true for most protein folds, the complexity can be reduced to approximately $O(n^2)$. However, SCRFs are still prohibitively expensive since the final complexity are multiplied by the number of iterations in an iterative search algorithm, which could be tens of thousands (see discussion in §4). In addition, the complexity will increase (exponentially) with the size of the cliques and indeterministic state transitions, which prevents it from large scale applications.

3. Chain graph model for protein fold recognition

In order to accurately predict the protein folds with structural repeats, it is crucial to consider the following two properties: 1) the structural motifs in each repeat have certain pleating and hydrogen bonding patterns that are well conserved across the superfamilies

and families; 2) the side-chain interactions between the neighboring motifs or insertions in 3-D are critical determinants of the stability of the structures (Kobe & Deisenhofer, 1994; Yoder & Jurnak, 1995; Kreisberg et al., 2000). Therefore, it is important for a model to be able to identify the sequence motifs reflecting the structural conservation, and at the same time consider the long-range interactions between structural elements. The SCRF model described above is not only prohibitively expensive computationally, but also lacks the device to incorporate sequence motif information. In this paper, we propose a chain graph model that makes use of both the undirected SCRFs and the directed sequence motif models as building blocks, and integrate them via a directed network. In this way, our model is able to capture the long-range interactions between structural repeats without computing a global normalizer required in SCRF.

3.1. Chain graph model

A *chain graph* is a graph consisting of both directed and undirected arcs associated with probabilistic semantics. It possesses the properties of both the Markov random fields (i.e., allowing potential-based local marginals that encode constraints rather than causal dependencies) and the Bayesian networks (i.e., not having a hard-to-compute global partition function for normalization and allowing causal integration of subgraphs that can be either directed or undirected) (Lauritzen & Wermuth, 1989). A chain graph can be represented as a combination of conditional networks. Formally, a chain graph over the variable set $\mathbf{V}$ that forms multiple subgraphs $\mathcal{U}$ can be represented by the following factored form: $P(\mathbf{V}) = \prod_{u\in\mathcal{U}} P(u|\text{parents}(u))$, where parents$(u)$ denotes the union of the parents for every variable in u. $P(u|\text{parents}(u))$ can be defined as a conditional directed or undirected graph (Buntine, 1995), which only needs to be *locally normalized*.

Back to the protein structure graph, we propose a *hierarchical segmentation* for a protein sequence. On the top level, we define an *envelope* Ξ_i, as a subgraph that corresponds to one repeat region in the fold containing both motifs and insertions or the null regions outside the protein fold. It can be viewed as a mega node in a chain graph defined on the entire protein sequence and its segmentation (Fig.2 (B)). Analogous to the SCRF model, let M denote the number of envelopes in the sequence, $\mathbf{T} = \{T_1, \ldots, T_M\}$ where $T_i \in \{\text{repeat, non-repeat}\}$ denotes the structural label of the i^{th} envelope. On the lower level, we decompose each envelope as a regular arrangement of several motifs and insertions, which can be modeled using one SCRF model. Let Ξ_i denote the internal segmentation of the i^{th} envelope (determined by the local SCRF), i.e. $\Xi_i = \{M_{(i)}, \mathbf{S}_{(i)}, \mathbf{T}_{(i)}\}$. Following the notational convention in the previous section, we use $W_{i,j}$ to represent a segment-specific clique *within* envelope i that completely determines the configuration of the j^{th} segment in the i^{th} envelope. To capture the influence of neighboring repeats, we also introduce a motif indicator Q_i for each top-level repeat i, which signals the presence or absence of sequence motifs therein, based on the sequence distribution profiles estimated from previous repeat. Putting everything together, we arrive at a chain graph depicted in Fig.2 (B).

Given a sequence $\mathbf{x}$, the value assignments of $\mathbf{W} = \{M, \{\Xi_i\}, \mathbf{T}\}$ in the chain graph G defines a hierarchical segmentation of the sequence as follows:

$$P(\mathbf{W}|\mathbf{x}) = P(M, \{\Xi_i\}, \mathbf{T}|\mathbf{x}) = \quad (3)$$
$$P(M)\prod_{i=1}^{M} P(T_i|\mathbf{x}, T_{i-1}, \Xi_{i-1})P(\Xi_i|\mathbf{x}, T_i, T_{i-1}, \Xi_{i-1}).$$

$P(M)$ is the prior distribution of the number of repeats in one protein and for simplicity a uniform prior is assumed. $P(T_i|\mathbf{x}, T_{i-1}, \Xi_{i-1})$ is the state transition probability and we use the structural motif as an indicator for the existence of a new repeat, i.e.:

$$P(T_i|\mathbf{x}, T_{i-1}, \Xi_{i-1}) = \sum_{Q_i\in\{0,1\}} P(T_i|Q_i)P(Q_i|\mathbf{x}, T_{i-1}, \Xi_{i-1}), \quad (4)$$

where Q_i is binary indicator denoting whether or not there exists a motif in the i^{th} envelope and $P(Q_i|\mathbf{x}, T_{i-1}, \Xi_{i-1})$ is computed using a profile mixture model described in §3.2. For the third term, we define the conditional probability using SCRF, i.e.

$$P(\Xi_i|\mathbf{x}, T_i, T_{i-1}, \Xi_{i-1})$$
$$= \frac{1}{Z_i}\exp\Big(\sum_{j=1}^{M_{(i)}}\sum_{k=1}^{K}\lambda_k f_k(\mathbf{x}, W_{i,j}, W_{\pi_{i,j}})\Big), \quad (5)$$

where Z_i is the *local normalizer* over the possible configurations of Ξ_i (instead of all envelopes), and $W_{\pi_{i,j}}$ is the spatial predecessor of $W_{i,j}$ defined by long-range interaction arcs. Similarly, parameters λ can be estimated by optimizing the regularized negative log-loss,

$$L_\lambda = \sum_{i=1}^{M}\sum_{j=1}^{M_{(i)}}\sum_{k=1}^{K}\lambda_k f_k(\mathbf{x}, w_{i,j}, w_{\pi_{i,j}}) - \log Z_i + \frac{\|\lambda\|^2}{2\sigma^2},$$

where the last term is a Gaussian prior over the parameters as a smoothing term.

Given a testing sequence, the optimal segmentation/labeling of the protein corresponds to state configuration with maximal conditional probability under our chain graph. Exploiting the chain structure induced by structural repeats and long range interactions, we propose a greedy search algorithm following similar idea as Viterbi algorithm. Define $\delta(s, t)$

as the highest score that the ending envelope are in state t given the observation $x_1 x_2 \dots x_s$, and $\varphi(s,t) = \{m, \mathbf{S}, \mathbf{T}\}$ is the corresponding "argmax" segmentation of the envelope. Then the recursive step is

$$\delta(r,t) = \max_{r',t',\xi} \delta(r',t') P(T=t|\mathbf{x}, t', \varphi(r',t')) P(\Xi = \xi|\mathbf{x}, t, t', \varphi(r',t')), \quad (6)$$

and $\varphi(s,t)$ equals to ξ that maximizes the eq(6).

To summarize, using a chain graph model, we can effectively identify motifs based on their structural conservation and at the same time take into account the long-range interactions between repeat units. In addition, a chain graph also reduces the computational costs by using local normalization. Since most side-chain interactions take effect within a small range in 3-D space, our model can be seen as a reasonable approximation for a global models as SCRF. For most protein folds, in which the length of one segment is much smaller than n or fixed, the complexity of our algorithm can be bounded by $O(nI)$, where I is the number of iterations in iterative searching algorithms.

3.2. Mixture profile model for structural motif detection

A commonly used representation for motif-finding is the position weight matrix (PWM), which records the relative frequency (or a related score) of each amino acid type at every position of a motif (Bailey & Elkan, 1994). Statistically, a PWM defines a product of multiple independent multinomial models over the observed instances of a motif.

An important observation in our task is that the motif instances close in three-dimension are more similar than those from distant locations or from different sequences. In addition, the residues with the side-chain pointing to the core are more conserved than those pointing outward. To capture these properties of structural motifs, a mixture PWM is proposed, which consists of a position-specific multinomial θ_j for the motif shared by all the proteins, and a sequence-specific multinomial $\theta_i^{(0)}$ for the background. Furthermore we define binary random variables $\mathbf{R} = \{R_{ij}\}$, where $R_{ij} = 1$ means that the j^{th} position in the i^{th} protein is generated by model θ_j and otherwise by model $\theta_i^{(0)}$. We assume that R_{ij} follows a Bernoulli distribution with parameter ρ_d, where d is the side-chain pointing directions (inward or outward) at position j. The parameters in the model can be learned using the EM algorithm straightforwardly. To calculate $P(Q_i|\mathbf{x}, T_{i-1}, \Xi_{i-1})$ in Eq (4), we do an online updating of $\theta^{(0)}$ and ρ using the motif instances defined by envelope (Ξ_{i-1}), then calculate the posterior as the probability that the sequence in Ξ_i in generated from the motif model θ divided by the likelihood define by the mixture.

Notice that the motif model described above is built specifically to capture the effects of neighboring motif instances, which is based on biological insights of the structures. So the motifs we learned are *site- and sequence-sensitive* and are different from the context-free motif profiles in databases, such as PROSITE and I-site (Bourne & Weissig, 2003).

4. Experimental Results

In our experiments, we test our algorithm on two important protein folds in β-class, i.e. the right-handed β-helices and leucine-rich repeats. We choose these two folds specifically because they are complex enough to represent the difficulties of the task, and well documented due to their important functions.

4.1. Experiment setup

We followed the setup described in (Bradley et al., 2001). A PDB-minus dataset was constructed from the PDB protein sequences (July 2004 version) (Berman et al., 2000) with less than 25% similarity to each other and no shorter than 40 residues. By removing the β-helix proteins (or LLR proteins) from it, the PDB-minus dataset can be used as the negative set for our validation. A leave-family-out cross-validation was performed, that is, for each cross, positive proteins in one SCOP family (see Table 1&2) are placed in the test set while the remainder are placed in the training set. Similarly, the PDB-minus set was also partitioned into the same proportion and for each cross we use one subset as testing data and the rest as training data. Since the ratio of negative examples to positive examples is very large, we subsample only 15 negative sequences that are most similar to the positive examples in sequence identity in order to find a better decision boundary.

We define two types of features for fold recognition. The first type is *Node features* covering the properties of an individual segment:

a *Regular expression template*: Based on the side-chain alternating patterns in the structurally conserved regions, a regular expression template is generated for β-helices as ΦXΦXXΨXΦX, where Φ matches any of the hydrophobic residues as {A, F, I, L, M, V, W, Y}, Ψ matches any residue except the ionisable ones {D, E, R, K}, and X is a wild card (Bradley et al., 2001). Similarly, the template for LLR is XXXLXXLX[LV]XXXXX. We define feature function $f_{RST}(x, w_i)$, which equals to 1 if the sequence in segment w_i matches the template, and 0 otherwise.

b *Probabilistic HMM profiles*: A probabilistic motif profile is built using HMMER (Durbin et al., 1998) to detect the structurally conserved regions as in (a). We define feature $f_{HMM}(x, w_i)$ as the alignment score of segment w_i against the profile.

Table 1. 0.980.3 Scores and rank for the known right-handed β-helices by HMMER, Threader, BetaWrap, SCRFs and chain graph model(CGM). 1: the scores and rank from BetaWrap are taken from [3] except 1ktw and 1ea0; The result of sequence-based HMMs (unlisted due to space limit) is much worse than struct-base HMMs.

SCOP Family	PDB-ID	Struct-based HMMs		Threader	BetaWrap[1]		SCRFs		CGM	
		Bit score	Rank	Rank	Wrap-score	Rank	ρ-score	Rank	ρ-score	Rank
P.69 pertactin	1DAB	-73.6	3	24	-17.84	1	10.17	1	31.69	1
Chondroitinase B	1DBG	-64.6	5	47	-19.55	1	13.15	1	34.89	1
Glutamate synthase	1EA0	-85.7	65	N/A	-24.87	N/A	6.21	1	29.04	1
Pectin methylesterase	1QJV	-72.8	11	266	-20.74	1	6.12	1	22.69	1
P22 tailspike	1TYU	-78.8	30	2	-20.46	1	6.71	1	20.59	1
Iota-carrageenase	1KTW	-81.9	17	10	-23.4	N/A	8.07	1	16.06	1
Pectate lyase	1AIR	-37.1	2	45	-16.02	1	16.64	1	22.87	2
	1BN8	180.3	1	76	-18.42	3	13.28	2	28.98	1
	1EE6	-170.8	852	228	-16.44	2	10.84	3	15.16	3
Pectin lyase	1IDj	-78.1	14	6	-17.99	2	15.01	2	17.50	2
	1QCX	-83.5	28	6	-17.09	1	16.43	1	20.67	1
Galacturonase	1BHE	-91.5	18	18	-18.80	1	20.11	3	28.98	1
	1CZF	-98.4	43	5	-19.32	2	40.37	1	24.68	3
	1RMG	-78.3	3	27	-20.12	3	23.93	2	27.37	2

c *Secondary structure prediction scores*: The state-of-art method of secondary structure prediction can achieve an average accuracy of 76 - 78%. It can provide fairly good results on α-helix and coils, which help to locate the insertions. We define feature function $f_{ssH}(x, w_i)$, $f_{ssE}(x, w_i)$ and $f_{ssC}(x, w_i)$ as the average of the predicted scores over all positions in segment w_i, for helix, sheet and coil respectively by PSIPRED (Jones, 1999).

d *Segment length*: $f_L(x, w_i) = (l - \mu)^2/\sigma^2$, where l is the segment length, μ and σ^2 are the mean and variance of the segment length in state T_i.

The second type of features are the *Inter-node features* capturing the potential long-range interactions between adjacent motifs in 3-D:

a *Side chain alignment scores*: It is suggested that the alignment scores of residue pairs in β-sheets are very discriminative features to identify long-range interactions between β-strands. A possible alignment scores is the conditional probability that a residue A_i aligns with residue A_j given their *side-chain orientation* relative to the structural core (Bradley et al., 2001). Following this idea, we define a feature $f_{SAS}(x, w_i, w_{\pi_i})$ as the weighted sum of the side chain alignment scores for w_i given w_{π_i} (see (Bradley et al., 2001) for full discussion).

b *Parallel β-sheet alignment scores*: Another aspect of the alignment scores is the different preferences between parallel and anti-parallel β-sheets. A "pairwise information values" is defined for a residue A_i given the residue A_j on the pairing parallel (or anti-parallel) strand within an offsets δ (Steward & Thornton, 2002). The alignment score for two segments $f_{PAS}(x, w_i, w_{\pi_i})$ is the sum of the pairwise information values over all the residues with an offset of no more than 2.

c *Distance between adjacent s-B23 segments*: We define the feature as the normalized length, i.e. $f_{DIS}(x, w_i, w_{\pi_i}) = (d - \mu')^2/\sigma'^2$, where d is the distance between w_i and w_{π_i}, μ' is the mean and σ'^2 is the variance.

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 BetaWrap, the state-of-art algorithm for predicting β-helices, THREADER, a threading algorithm and HMMER, a general motif detection algorithm using HMMs. 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).

4.2. β-helices

The β-helix fold is an elongated helix-like structure whose repeat units are composed of three parallel β-strands, namely B_1, B_2 and B_3 strand (see Fig.1). The regions connecting these strands are called T_1, T_2 and T_3 turn respectively. In particular, T_2 turn is structurally conserved as a unique two-residue turn which forms an angle of approximate 120$^\diamond$ between the B_2 and B_3 strands. Therefore we define 2 structural *motifs* for the β-helix fold, one is the union of B_2, T_2 and B_3 with 9 residues in total, the other is B_1 strand with 4 residues. The length of the *insertions* connecting the motifs varies from 1 to 80 residues.

There currently exist 14 protein sequences with β-helix whose crystal structures have been known. Those proteins belong to 9 different SCOP families (Murzin et al., 1995) (Table 1). Computationally, it is very difficult to detect the β-helix fold because the proteins with this fold are less than 25% similar in sequence identity, which is the "twilight zone" for sequence-based methods, such as PSIBLAST or HMMs, and there involve the long-range interactions. The state-of-art method is BetaWrap, which is a heuristic methods specifically designed for the β-helix (Bradley et al., 2001). The algorithm works by identifying all potential motifs in the sequence and "wrapping" them to see if they can form a stable structures.

Table 1 shows the output scores by different meth-

Table 2. Scores and rank for the known right-handed Leucine-rich repeats (LLR) by HMMER, Threader and chain graph model (CGM). For CGM, ρ-score = 0 for all non-LLR proteins.

SCOP Family	PDB-ID	ClustalW+HMMs		Struct-based HMMs		Threader	CGM	
		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 A/CDK2-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

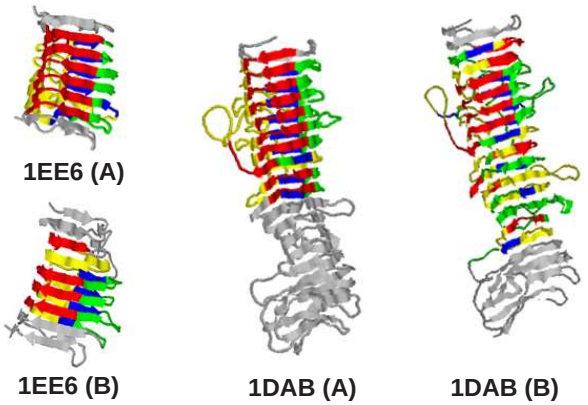

Figure 3. Segmentation for protein 1EE6 and 1DAB by SCRFs(A) and chain graph model (B). Red: B2-T2-B3 motif; blue: B1 motif; green and yellow: insertions.

ods and the relative rank for the β-helix proteins in the cross-family validation. From the results, we can see that the both SCRFs and chain graph 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 lower than the scores of other β-helices and some non β-helix proteins, which indicates that BetaWrap is over-trained. As expected, HMMER performs worse than other methods even using the structural alignments.

Our algorithm also demonstrates success in locating each repeat in the known β-helix proteins. Fig.3 shows the segmentation results for 1EE6 and 1DAB respectively. From the results, we can see: for 1EE6 SCRFs can locate two more repeats accurately than the chain graph model; however, our model is able to span the repeats over the whole area of the true fold for 1DAB while SCRFs can only locate part of them. We can see that there are strength and weakness for both methods in terms of segmentation results. On the other hand, since the computational complexity for chain graph model is only $O(N)$, the real running time of our model (approx. 2.5h) is more than 50 times faster than that of SCRFs (approximately 140h).

4.3. Leucine-rich repeats

The leucine-rich repeats are solenoid-like regular arrangement of β-strand and an α-helix of variable lengths, connected by coils (Fig.1). Based on its structural characteristics, we define the *motif* for LLR as the β-strand and short loops on two sides, resulting 14 residues in total. The *insertions*, which consist of the α-helix and some loops, have a length from 6 to 29 (since longer insertions will destroy the stability of the structures). There are 41 LLR proteins with known structure in PDB, covering 2 super-families and 11 families in SCOP. The LLR fold is relatively easy to detect due to its conserved motif with many leucines in the sequence and relatively 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 2 lists the output scores by different methods and the rank for the LLR proteins. We can see that LLR is generally easier to identify than the β-helices. The chain graph model also performs much better than other methods by ranking all LLR proteins higher than non-LLR proteins. In addition, the predicted segmentation by our model is close to prefect match for most LLR proteins. Some examples are shown in Fig.4.

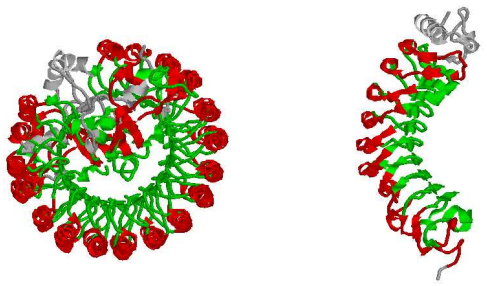

Figure 4. Segmentation for protein 1OGQ and 1A4Y by chain graph model. Green: motif; red: insertions.

5. Conclusion

In this paper, we introduce a chain graph model to identify an important type of complex protein folds, i.e. those with structural repeats. Our model makes use of both the undirected SCRFs to deal with long-range interactions and the directed sequence motif models as building blocks. It integrates the two parts gracefully via a directed network under the framework of chain graph models. The experimental results on β-helices and LLRs show that our model performs significantly better than the previously proposed methods in predicting the membership of protein folds. In addition, it is much more efficient than the SCRFs model for general fold recognition.

It is worth noting that although our discussion has focused on applying the chain graph technique to protein fold recognition, the long-range interactions/dependencies are common phenomena in many applications, such as machine translation or information extraction. We anticipate that the approach presented here can be straightforwardly extended for recognizing more challenging protein folds and for other prediction tasks in IR and NLP.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0225656 and Pennsylvania TSF grant. We thank anonymous reviewers for their valuable suggestions.

References

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, *25*, 3389–402.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. of ISMB'94*.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., & Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, *28*, 235–42.

Bourne, P. E., & Weissig, H. (2003). *Structural bioinformatics: Methods of biochemical analysis*. Wiley-Liss.

Bradley, P., Cowen, L., Menke, M., King, J., & Berger, B. (2001). Predicting the beta-helix fold from protein sequence data. *Proceedings of ACM RECOMB'01*.

Buntine, W. L. (1995). Chain graphs for learning. *Uncertainty in Artificial Intelligence* (pp. 46–54).

Chu, W., Ghahramani, Z., & Wild, D. L. (2004). A graphical model for protein secondary structure prediction. *Proc.of ICML'04*.

Delcher, A., Kasif, S., Goldberg, H., & Xsu, W. (1993). Protein secondary-structure modeling with probabilistic networks. *Proc. of ISMB'93* (pp. 109–117).

Ding, C., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics.*, *17*, 349–58.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.

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

Hammersley, J., & Clifford, P. (1971). *Markov fields on finite graphs and lattices*. Unpublished manuscript.

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

Kobe, B., & Deisenhofer, J. (1994). The leucine-rich repeat: a versatile binding motif. *Trends Biochem Sci.*, *10*, 415–21.

Kreisberg, J., Betts, S., & King, J. (2000). Beta-helix core packing within the triple-stranded oligomerization domain of the p22 tailspike. *Protein Sci.*, *9*, 2338–43.

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

Lafferty, J., Zhu, X., & Liu, Y. (2004). Kernel conditional random fields: representation and clique selection. *Proc.of International Conference on Machine Learning (ICML-04)*.

Lauritzen, S., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, *17*, 31–57.

Liu, Y., Carbonell, J., Weigele, P., & Gopalakrishnan, V. (2005). Segmentation conditional random fields (SCRFs): A new approach for protein fold recognition. *Proc. of ACM RECOMB'05*.

Minka, T. P. (2001). Algorithms for maximum-likelihood logistic regression. *CMU Statistics Tech Report 758*.

Murzin, A., Brenner, S., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.*, *247*, 536–40.

Steward, R., & Thornton, J. (2002). Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins.*, *48*, 178–91.

Thompson, J., Higgins, D., & 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 Research*, *22*, 4673–80.

Yoder, M., & Jurnak, F. (1995). Protein motifs. 3. the parallel beta helix and other coiled folds. *FASEB J.*, *9*, 335–42.

Yoder, M., Keen, N., & Jurnak, F. (1993). New domain motif: the structure of pectate lyase c, a secreted plant virulence factor. *Science*, *260*, 1503–7.