

Prediction of parallel and anti-parallel beta-sheets using Conditional Random Fields

Yan Liu¹, Jaime Carbonell¹, Judith Klein-Seetharaman^{1,2}, Vanathi Gopalakrishnan²

1. School of Computer Science, Carnegie Mellon University
2. Center for Biomedical Informatics, University of Pittsburgh

Abstract

In this paper, we present a new algorithm for parallel and anti-parallel beta-sheet prediction using conditional random fields. In recent years, various approaches have been proposed to capture the long-range interactions of beta-sheets. However, most of them are not very successful: either the learning models are not general enough to capture the non-local information, or the features they used do not contain the information, for example, the window based profiles. Our new method has the advantages over previous methods in two aspects: (1) It takes into account both the local information and long-range interaction information (2) The condition random fields are powerful models that are able to capture long-range interaction features. The experimental results show that our algorithm performs significantly better than the state-of-art secondary structure prediction methods.

1. Introduction

Protein secondary structure prediction involves the projection of primary protein sequences onto a string of secondary assignments, such as helix, sheet or coil, for each residue. It is widely believed that secondary structure prediction can contribute valuable information to discerning how proteins fold in three-dimensions. In Recent years, the performance of protein secondary structure prediction has improved to as high as 78% in accuracy in general and 80-85% for predicting helix and coils. The major bottleneck lies in predicting beta-sheets, which involves long-range interactions between regions of the protein chain that are not necessarily consecutive in the primary sequence. Therefore designing an effective algorithm for detecting beta-sheets will help to improve the performance for secondary structure predictions and is the main focus of this paper.

As we know, most beta-sheet structures consist of either parallel or anti-parallel strands, or mixed sheets (both parallel and anti-parallel). In some reduced 3-class definitions for protein secondary structures, the beta-bridge, with only one amino acid on each strand, is also counted as sheets. In this paper, we only focus on the prediction of beta-sheets using the first definition. We are interested in answering the following questions: given a protein sequence, which residues belong to beta-sheets class, in particular we try to predict the strand pairing in parallel or anti-parallel strands.

Various approaches have been proposed to capture the long-range interactions of beta-sheets. Mamitsuka & Abe used stochastic tree grammars for beta-sheet detection [Mamitsuka & Abe, 1994]. However, their experimental results are based on only two protein sequences. Baldi et al. computed a number of beta-sheet and beta-strand

statistics, including the first order statistics, i.e. frequencies for each amino acids in background and in the beta-sheet structures, second order statistics, i.e. the pairings of amino acids, and length distribution [Baldi et al, 2000]. These statistics are then fed into a bi-directional recurrent neural networks(BRNN). Their experiments on 836 protein sequences shows an accuracy of 84%. Although the numbers are very encouraging, the data sets they used are very unbalanced so that the negative residues cover around 90% of the data. Recently, Steward & Thornton developed to a set of tables with pair-information values for each amino acids using information theory approach [Steward & Thornton, 2002]. Using the tables, they can correctly aligned two beta-strands 45% and 48% of anti-parallel and parallel alignments.

Most of previous attempts to solve the non-local interactions are not very successful for two reasons: either the learning models are not general enough to capture the non-local information, or the features they used do not contain the information, for example, the window based profiles. To solve the first problems, we proposed to use the conditional random fields (CRFs). CRFs, as a graphical chain model, is similar to hidden markov models (HMMs), but it is more general than HMMs in two aspects: (1) the optimization function in CRFs is convex, so the solution is global optimal. (2) CRFs is able to take into any features from the observation sequences, which enable CRFs to deal with long-range interaction problem. To solve the second problem, we propose to use the both the local information, i.e. the profiles from PSI-blast and the non-local pairing information values by Steward & Thornton.

The rest of the paper is organized as follows: in section 2, we discuss the problem definition in detail; Then the features, i.e. the statistics we used for our prediction are described; in section 4, we introduce our algorithm and show the experimental results; finally, we give discussion and shed light on how to use our results to help improve secondary structure predictions.

2. Problem Definition

2.1 Parallel and anti-parallel beta-sheets

The beta-sheets consists of either parallel or anti-parallel strands, or mixed parallel and anti-parallel strands. A beta-strand can have 1 or 2 pairing strands, i.e. each individual amino acid can have one or two pairing residues. However, how far apart the amino acids can find on another to form a beta-sheets is still poorly understood. Sometimes, the beta-strand partner are even located on a different protein chain. All of these properties of make it difficult to predict the beta-sheets accurately.

In some reduced 3-class definitions for protein secondary structures, the beta-bridge, with only one amino acid on each strand, is also counted as sheets. We exclude the discussion for beta-bridge prediction since they have different properties and possibly different formation reasons from parallel and anti-parallel beta-sheets.

Without loss of generosity, we formularize our target problem as follows: given a protein sequence $R_1R_2R_3\dots R_N$, and a label set $\{N, E\}$, where N refers to NOT beta-sheets

and E refers to beta-sheets, the output is a sequence of labels $L_1L_2L_3\dots L_N$, where $L_i = N, E$.

2.2 Data collection and label assignment

In our experiments, we used two datasets: one is the RS126 dataset, on which many previous secondary structure prediction methods were developed and tested. It is a non-homologous dataset and was developed by Rost & Sander: no two proteins of 126 protein chains share more than 25% sequence identity over a length of more than 80 residues [Rost B & Sander C, 1993]. The other is the CB513 dataset, constructed by Cuff & Barton. It consists of 513 non-homologous protein chains which have an SD score, i.e. Z score for comparison of the native sequences given by $(V-x)/\sigma$, of less than 5 [Cuff & Barton, 1999]. The two datasets can be downloaded from the web <http://barton.ebi.ac.uk/>.

The definition we followed in our experiments is the DSSP definition [Kabsch & Sander, 1983], which is based on hydrogen bonding patterns and geometrical constraints. In DSSP, a beta-sheet is beta-sheet ("E") is defined when it forms two H-bonds in the sheet or is surrounded by two H-bonds in the sheet. The minimal sheet length is 2. A standard example of how the pair assignments are made is shown in figure-1.

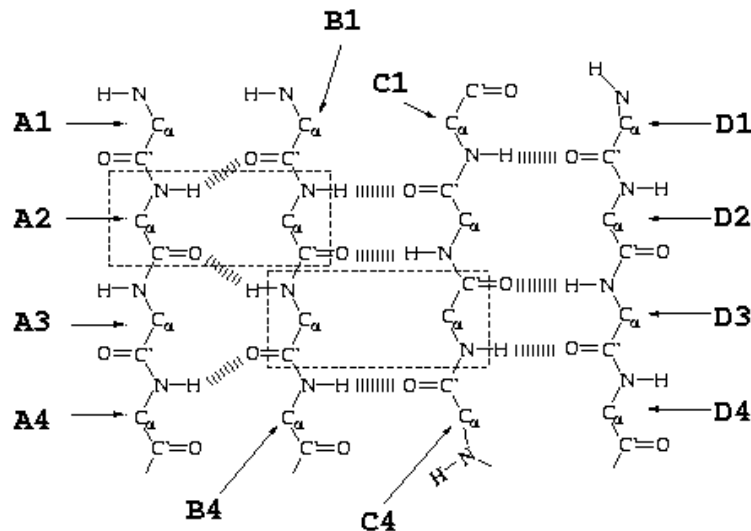


Figure -1. The assignment criterion for sheet partners. Notice that the in the special case of beta-bulge, B3 and C3 are also counted as partner even though there are no H-bonds in-between.

3. Useful features for prediction

3.1 Pair-information Values for Parallel and Anti-parallel strands

The pair-information values for beta-sheets are used to predict the correct alignment of two beta-strands [Steward & Thornton, 1999]. The score for two sequences consists of an intra-strand score (S) and inter-strand score (P). Particularly, for a pair of residue R_i

and R_j at position i and j on adjacent strands, the score consists of the self-information (S_i and S_j), i.e. the information as to whether R_i and R_j in a beta-sheet, and pair-information ($P_{i,j}$ and $P_{j,i}$), i.e. the information contributed by interacting residues on adjacent strand.

Self Information Score (S_i) for a residue R_i at position i in a parallel (or anti-parallel) beta-strand is defined as $S_i = \log(f_{\beta,R_i} / f_{N\beta,R_i}) + \log(f_{N\beta} / f_{\beta})$. Seen from probability point of view, $S_i = \log(P(R_i | \beta) / P(R_i | N\beta))$.

Pair-information ($P_{i,j}$) is the information for residue R_i , held by a residue R_j on an adjacent beta-strand. The pair-information for residue R_i is calculated for all the residues R_{j+m} offset by m from position j . It is defined as follows: $P_{i,j+m} = \log(f_{\beta,R_{j+m}|R_i} / f_{N\beta,R_{j+m}|R_i}) + \log(f_{N\beta,R_i} / f_{\beta,R_i})$. Seen from probability point of view, $P_{i,j+m} = \log(P(R_{j+m} | R_i, \beta) / P(R_{j+m} | R_i, N\beta))$.

Total score for a particular alignment is the sum of, for each position on both strands, the self-information value given by the residue type on the same beta-strand, and the pair-information value given by residue types on adjacent strand. The definition is as follows:

$TotalScore = \sum_{i,j}^{R_i,R_j} (S_i + S_j + \sum_{m=-2}^{m=2} (P_{i,j+m} + P_{j,i+m}))$. Seen from the probability point

of view, we have

$$\begin{aligned} S_i + \sum_{m=-2}^{m=2} P_{i,j+m} &= \log(P(R_i | \beta) / P(R_i | N\beta)) + \sum_{m=-2}^{m=2} \log(P(R_{j+m} | R_i, \beta) / P(R_{j+m} | R_i, N\beta)) \\ &= \log(P(R_i, R_{j-2}, R_{j-1}, R_j, R_{j+1}, R_{j+2} | \beta) / P(R_i, R_{j-2}, R_{j-1}, R_j, R_{j+1}, R_{j+2} | N\beta)). \end{aligned}$$

Similarly,

$$S_j + \sum_{m=-2}^{m=2} P_{j,i+m} = \log(P(R_j, R_{i-2}, R_{i-1}, R_i, R_{i+1}, R_{i+2} | \beta) / P(R_j, R_{i-2}, R_{i-1}, R_i, R_{i+1}, R_{i+2} | N\beta)).$$

Therefore the total score can be seen as the log ratio of the joint probability of the paired beta-strands.

The score tables were first developed to predict the alignment of two beta-strands given the knowledge of the length of the beta-ladder, the direction of the strands and the sequence of one strand. However, how to use the pair information values for parallel and anti-parallel strand prediction has not been explored and in section 3.3 we will discuss in detail our methods of using this pair information.

3.2 PSI-blast profile

In recent years, one of the major improvement in protein secondary structure prediction is due to the use of position-specific scoring matrices (PSSM), i.e. the PSI-blast profile [Jones, 1999]. PSSM was initially used for searching sequences in the database by PSI-blast. It is an $20 \times N$ matrix A , where N is the length of the sequence. $A(i, j)$ is the

estimation of the probabilities of amino acid i occur at the position j . It has been proved that PSSM is a more sensitive scoring matrices and can relatively precisely detect the boundaries of important motifs.

In order to use effectively of the sensitive PSI-blast profile, a custom sequence data bank was constructed. The databank contains all the protein sequences in PDB and Swissprot, which are then filtered with SEG program to remove regions with very low information content, transmembrane segments, and coiled-coil regions. In our experiments, the position specific scoring matrix generated from the PSI-BLAST after three iterations are fed into SVM classifiers. The matrix elements are scaled by using linear transformation proposed in [Kim & Park, 2003]:

$$f(x) = \begin{cases} 0 & \text{if } (x \leq -5) \\ 0.5 + 0.1x & \text{if } (-5 < x < 5) \\ 1.0 & \text{if } (x \geq 5) \end{cases}$$

where x is the value from the PSSM.

3.3 Long-range interaction feature construction

It has long been known that the major bottleneck for protein secondary structure prediction is the beta-sheets, which involve long-range interactions. In this paper, we present a new kind of features that tries to combine the local information with the pairing values to provide the relatively reliable long-range interaction features.

The basic idea of our method is: given the predictions from the window-based method using PSI-blast profiles, we can have certain confidence that some of the residues R_x belong to beta-sheets. Then we can search in the whole sequence to identify the corresponding paired beta-strand in the long range. If R_x has a potential paired beta-strand R_y with a high score, then we can successfully identify R_y as beta-sheets. On the other hand, if there are no suitable beta-strand in the sequence, then R_x may not belong to beta-sheet at all. In this way, we can reinforce the correct prediction and at the same time adjust the incorrect predictions. The detailed algorithm is shown below:

Step1: Given a protein sequence $X_1X_2\dots X_N$, extract the Position specific score matrices A using PSI-blast.

Step 2: Input PSSM A as features to SVM classifiers. The output is a score matrix S , which is a $3*N$ matrix since there are 3 classes, N is the length of the protein sequence.

Step 3: Build an $N*N$ paired scoring matrix K_{para} , K_{anti} for the sequence using parallel and anti-parallel paired-value separately. $K(i, j)$ is the paired score between residue X_i and residue X_j , defined as $K(i, j) = S_i + S_j + \sum_{m=-2}^m (P_{i,j+m} + P_{j,i+m})$. We normalize the score by $(K-k)/\sigma$, where k is the mean and σ is the variance.

For the following steps, we maintain a active list L. For step 4 and 5, we only discuss the case for parallel strands prediction, the anti-parallel cases can be developed similarly.

Step 4: For each of the residues R_i in the sequence: if R_i is predicted as belong to E, then find all the possible pair residues R_j with the score $K(i, j)$ above some threshold T1. R_i and R_j must be 7 residues since local interactions can be captured the by Step-2. Then we extend the pair (R_i, R_j) to neighbor residues with the constraint that $K(i+/-m, j+/-m)$ is above threshold T1. The total score for the two strands will be the sum of the scores for each pair. We add the extended strands to the active list L if the total score is above some threshold T2.

Step 5: Sort the strands in the active list L in descending order of the corresponding scores. Filter the strands with common residues by leaving the strand with the highest scores and deleting all the others from the active list L.

Finally we get the active list L containing the possible strands involving non-local interactions.

4. Conditional random fields for Beta-sheet prediction

4.1 Conditional random fields

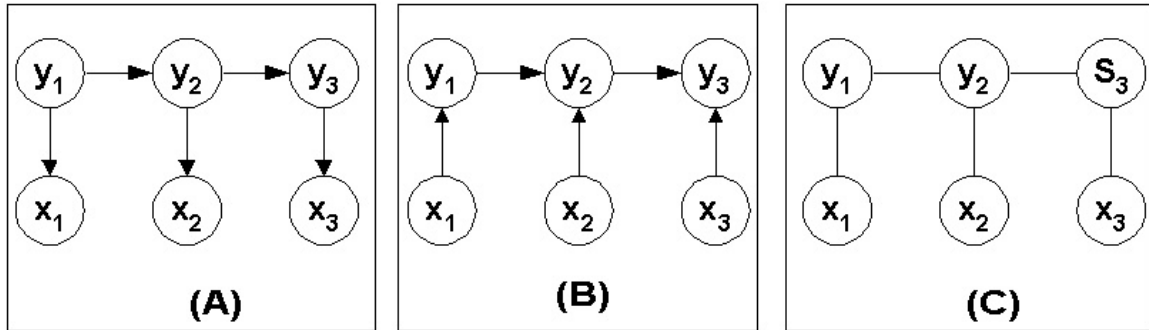


Figure-2 Graphical Structures of simple HMM(A), MEMM(B), and chain-structured CRF(C)

After getting the long-range interaction information, the next step is to discover where how to use them. Conditional Random Fields(CRFs, also known as Markov random fields) , proposed by Lafferty et al., are undirected graphical models to calculate the conditional likelihood $P(y|x)$ [Lafferty et al, 2001]. In chain-form, the CRFs graphical structure is shown in Figure -1. Unlike Hidden Markov Model(HMMs), CRFs are undirected graphical models.

By the Hammersely-Clifford theorem, CRFs define the conditional probability $P(y|x)$ to be proportional to the product of potential functions on the cliques of the graph,

$$P(y|x) = \prod_{c \in C(x,y)} \Phi_c(x_c, y_c) / Z_0,$$

where $\Phi_c(x_c, y_c)$ is the clique potential on the clique c ,

$C(x,y)$ is the set of cliques on the graph and Z_0 is the normalization factor over all y

values. Linear-chain CRFs obeys Markov properties, and the conditional probability $P(y | x)$ is defined as

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

where Z_0 is a normalization factor, namely $Z_0 = \sum_y \exp\left(\sum_{i=1}^N \sum_k \lambda_k f_k(y_{i-1}, y_i, x)\right)$, f_k

represents features that depend on y_i and on *any* properties of the input sequence. Therefore CRFs can support long-distance interactions. λ is learned via maximizing the conditional likelihood of the training data. Comparing with HMMs, the optimization function of CRFs are convex, so there is global optimal solution. Besides, CRFs can support long-range information while HMMs cannot.

The definition for the "forward values" $\alpha_i(Y)$ is slightly modified to be the probability of being in state Y at time i *given* the observation up to time i . The recursive step is as follows:

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

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

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

The backward procedure and the Viterbi algorithm can be derived accordingly.

4.2 Beta-sheet prediction

In this section, we will define the features for beta-sheet prediction using conditional random fields f_k .

In our method, there are two kinds of features involved: one is the prediction score from the SVM classifier S , i.e. the local information, the other is the active list L , i.e. the long-range information. To be exact, the features are defined as follows:

$$f_{\langle j,k \rangle}^{local}(y_{i-1}, y_i, x, i) = \begin{cases} S_i(k) & \text{if } (y_i = k) \\ 0 & \text{otherwise.} \end{cases} \quad f_{\langle j,k \rangle}^{non-local}(y_{i-1}, y_i, x, i) = \begin{cases} 1 & \text{if } (k = E) \\ 0 & \text{otherwise.} \end{cases}$$

In this way, we can capture both the local information and long-range interactions of beta-sheets.

5. Experimental Results

In our experiments, we use the SVM classifier with the PSI-blast profile features as baseline to compare our algorithm. Various performance measures are used to evaluate the prediction accuracy, including overall per-residue accuracy Q_3 , Matthew's correlation coefficients for structure type (C). The per-residue accuracy for each type of secondary structure (Q; Q^{pre}) was also calculated. Q^{rec} and Q^{pre} is distinguished as:

$$Q^{rec} = \frac{\# \text{ of residues correctly predicted in state I}}{\# \text{ of residues observed in state I}} \times 100$$

$$Q^{pre} = \frac{\# \text{ of residues correctly predicted in state I}}{\# \text{ of residues predicted in state I}} \times 100$$

Seven-fold cross-validation was used for both datasets, which is the same setting as reported in [Rost & Sander, 1993].

Table -1 lists the results of our new algorithm as compare to the state of art algorithm. From the results, we can see that our new algorithm is indeed good at predicting beta-strands, even better than the state of art methods for protein secondary structure prediction. Our algorithm is also better than compared with the results reported by [Baldi et al, 2000], which is 84% in accuracy considering that they used a different dataset.

	RS-126 dataset			
algorithm	Accuracy(Q3)	Q ^{rec}	Q ^{pre}	C
PSI_pred	87.48	60.30	75.87	60.10
Pairwise Information	76.56	55.10	50.34	N/A
New algorithm	88.20	63.12	74.98	N/A

Table-1 Comparison of our new algorithm and the SVM classifier using PSI-blast profile

6. Discussion and Conclusion

In this paper, we propose a new method for parallel and anti-parallel beta-sheet prediction using conditional random fields. This method overcomes the disadvantages of previous methods in two aspects (1) It takes into account both the local information and long-range interaction information (2) The condition random fields are powerful models that are able to capture long-range interaction features. The experimental results show that our algorithm performs significantly better than the state-of-art secondary structure prediction methods.

Currently the major bottleneck for secondary structure prediction lies in the beta-sheets. Recently Jones improved the prediction accuracy to 80% by using the neighboring information extracted from 3-D structures. Our algorithm is able to adapt to secondary structure prediction by using the pairing information of two beta-strands. Those paired residues should be close in 3-D structures. In this way, we can get the neighboring information without going to 3-D structure prediction. Therefore how to use the results of our algorithm to improve the overall performance of secondary structure prediction is one direction for our further research. In addition, we only consider the parallel and anti-parallel beta-sheets. There are a lot of other structures, such as beta-bridage and beta-helix. How to adapt our algorithm to effectively identify those structures is also an interesting next step.

Reference:

Baldi P, Pollastri G, Andersen CA, Brunak S (2000). Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:25-36.

Cuff JA, Barton GJ (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins.*, 34(1999) pp 508-519.

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

Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.* 2003 Aug;16(8):553-60.

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999 Sep 17;292(2):195-202.

Lafferty J, Pereira F, McCallum A(2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML'01)*, 2001.

Mamitsuka H, Abe N (1994). Predicting location and structure of beta-sheet regions using stochastic tree grammars. *Proc Int Conf Intell Syst Mol Biol.* 1994;2:276-84.

Rost B, Sander C (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol.*, 232(1993) pp 584-599.

Rost B, Sander C, Schneider R (1994). Redefining the goals of protein secondary structure prediction. *J Mol Biol.*, 235(1994) pp 13-26.

Ruczinski I, Kooperberg C, Bonneau R, Baker D (2002). Distributions of beta sheets in proteins with application to structure prediction. *Proteins.* 2002 Jul 1;48(1):85-97.

Steward RE, Thornton JM (2002). Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins.* 2002 Aug 1;48(2):178-91.

Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins.*, 34(1999) pp220-223.