Journal of Computer and Knowledge Engineering, Vol. 1, No. 1.
DOI: 10.22067/cke.v1i2.56005

β-sheet Topology Prediction Using Probability-based Integer Programming

Mahdie Eghdami*, Toktam Dehghani, Mahmoud Naghibzadeh

Abstract. β-sheet topology prediction is a major unresolved problem in modern computational biology. It is a challenging intermediate step toward the protein tertiary structure prediction. Different methods have been provided to deal with the problem of determining the β-sheet topology. Here, ab-initio probability-based methods called "*BetaProbe1*" and "*BetaProbe2*" are utilized to specify the β-sheet topology. In these methods, the stability and the frequency of β-strand pairwise interaction and β-sheet conformation are spotted. To predict more frequent interactions between β-strand pairs, besides pairwise alignment probability, the probability of occurring β-strand pairwise interaction is considered to compute the score of the interactions. Furthermore, to determine the β-strand pairwise alignment probability more accurately, a dynamic programming approach is utilized. In addition, the integer programming optimization is combined with the probabilities of β-strand pairwise interactions to determine the β-sheet topology. Moreover, the β-sheet conformation probability is considered to give better chances to more observed conformations for selection. Experimental results show that *BetaProbe1* and *BetaProbe2* significantly outperform the most recent integer programming-based method with respect to β-sheet topology prediction.

Keywords: β-sheet topology prediction; integer programming; dynamic programming; pairwise alignment;

1.Introduction

Proteins perform critical functions within the living organisms. Biologists believe that the functionality of proteins is determined by their tertiary structures. Therefore, it is important to specify the protein structure. Further, the conventional empirical methods to determine the structure of protein, namely, X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy are very costly, time-consuming, and sometimes impossible. In addition, now, from the 30 million proteins with known primary structures in the protein databases [1], only the tertiary structures of 30 thousand of them have been determined by experimental methods [2]. Therefore, there is a huge gap between the number of known primary structures and the number of determined tertiary structures. Hence, insufficiency of empirical methods leads to utilizing computational methods in protein structure prediction problem.

One of the most frequent elements in the protein structure is β-sheet which consists of separate sections known as β-strands. β-strands are typically six to eight amino acids long [3] that interact with amino acids of other β-strands and make paired β-strands (partners). The interaction between two β-strands can occur in two different forms (parallel or anti-parallel) depending on their orientation given by the position of the β-strands' N- and C-termini [4]. Each amino acid in a β-strand can make at most two hydrogen bonds with other ones in the paired stand. The interactions between the amino acid residues of the paired β-strands are known as a β-contact map.

β-sheets can be open or closed. Open β-sheets have two edge strands and they are the most common types of β-sheets. Fig. 1 shows an example of an open β-sheet type, where four β-strands interact. On the other hand, in the closed ones a circle is formed by a hydrogen bond between the first strand and the last one.

Fig. 1. Open β-sheet of a protein with PDB (Protein Data Bank) id 1NZ0D. β-strands that form the β-sheet are numbered in sequential order.

β-sheet topology prediction is regarded as one of the most important unresolved problems toward the tertiary structure prediction of proteins [5]. Correct prediction of β-sheet topology remains challenging because of hydrogen bond formations between linearly distant β-sheet residues [4]. Furthermore, the global covariations and constraints characteristic of β-sheet structures have not been well exploited [4].The β-sheet topology prediction provides valuable information for predicting protein three-dimensional structure [6], [7], designing new proteins and new drugs [8], [9] and determining folding pathways [10], [11].

Manuscript received May 16, 2016; accepted September 24, 2016.
Department of Computer Engineering, Engineering Faculty, Ferdowsi University of Mashhad, Mashhad, Iran.
*The corresponding author's e-mail is: eghdami.mahdie@mail.um.ac.ir

The main goal of predicting β-sheet topology from the protein's amino acids is to determine the organization of β-strands in the β-sheets. This includes identifying β-strand members of each β-sheet and describing β-sheets by specifying paired β-strands and their interaction types. Further, β-contact maps are determined in β-sheet structure prediction. Different methods have been proposed to address the problem of predicting β-sheet topology which will be described in the next section.

In this article, we present *BetaProbe1* [12] and *BetaProbe2*, ab-initio probability based methods for β-sheet topology prediction. The main advantage of the proposed methods as compared to the previous researches is that we make use of the fact that more frequent and more stable conformations should have greater chances of being selected. For this purpose, the score of an interaction between each two β-strands is computed considering both pairwise alignment probability and pairwise interaction probability. Moreover, in order to make more accurate alignments, the β-strand optimum pairwise alignment is found using a dynamic programming approach. Furthermore, combining integer optimization with the β-strand pairwise interaction probability improves the accuracy of the predicted interactions. In addition, using β-sheet conformation probability in the last step of *BetaProbe1* leads to predicting more frequent and more stable conformations.

In the rest of this paper, first, related studies are reviewed in Section 2. Then, the details of the proposed methods will be described in Section 3. Finally, the performances of the proposed methods are compared with the most recent integer programming-based β-sheet prediction method in Section 4.

2. Related Work

Most β-sheet topology prediction methods utilize contact maps and strands alignment. Any improvement in the accuracy of these fields leads to a higher accuracy in determining the architecture of β-sheets. In this section first the related works in these fields are introduced. Then, some β-sheet prediction methods are explained.

Specifying the protein contact map is the first step in determining its final structure. Mainly, a contact map is expressed by a two-dimensional matrix. For two amino acids r_i and r_j, if the value of the i-th row and the j-th column ($0 \leq$contact Map (i, j) ≤ 1) is closer to one then they are more likely to interact with each other in the final structure. In other words, the likelihood of their relationship in the final structure of proteins is higher. NNcon [13], DNcon [14], SVMcon [15] and Distill [16] can be mentioned as contact map prediction methods. CMAPpro [17], PSICOV[18] and PhyCMAP [19] are the most recent methods which include contact map prediction.

So far, methods with high accuracy and acceptable execution time have been suggested for the sequence alignment problem. Further, pairwise sequence alignment is the most common technique used in β-sheet prediction methods. The most usual approach to determine the best alignment between two strands is dynamic programming.

Many efforts have been made to address the problem of predicting β-sheet topology. These works can be divided into two major categories: homology-based methods and ab-initio methods. The homology-based methods such as SMURF [20], SMURFLite [21], and MRFy [22] use homological information of proteins for recognizing their topologies. On the other hand, ab-initio methods only consider amino acids' pairing potentials and statistical information. In this article, we concentrate on the ab-initio β-sheet topology prediction methods. They utilize different approaches such as statistical potentials[23], information theory[24], Bayesian models and exploration of entire search space[25], linear programming [5], [26], [27], hidden Markov models [28], and graph matching algorithms [4]. These approaches can be divided into two major categories[29]: in one category, all possible β-topologies are enumerated, and a score for each complete β-topology is computed. Then, the β-topology with the highest score is selected as the best one [7], [25]. In the other category, in order to predict the β-sheet topology of a protein, pseudo-energy is assigned to each pair of β-strands. Then the problem of determining the best β-topology is reduced to maximizing the strand-to-strand contact potentials of the protein [5], [4], [26], [27], [28], [30].

BetaPro [4] was the first method to take into consideration the global nature of β-sheet topologies. In this method, three stages are used to predict β-topologies. Jones [31] takes advantage of linear programming to predict the secondary structure of the protein and β-sheet topologies. In [27], BetaPro was combined with linear programming to predict β-sheet topologies. Also, Rajgaria et al. [30] presented a method to determine the tertiary structure of proteins. In this method, strand pairing scores and contact maps are computed using linear programming. BetaZa[25] is a Bayesian approach which was introduced for proteins up to six β-strands. The conformational features were modeled in a probabilistic framework. The model is a combination of prior knowledge about β-strand arrangements with pairing potentials between the strands amino acid. Also, to select the optimum β-sheet architecture, using some heuristics, the search space was reduced. A dynamic programming was used to determine the β-strands optimum pairwise alignment. In the proposed dynamic programing, any number of gaps were allowed. As a result of exploration approach of the entire search space, BetaZa has a high time complexity. BeST [5] and BCov [26] predict the β-sheet topology using integer programming. BCov determines the β-sheet topology in three steps: first, it computes the residue contact propensity using PSICOV[18]; then, it computes the score of each possible β-strand pairing. Finally, an integer programming optimization is used to determine the β-sheet topology by finding the best solution according to the constraints and the pairing scores. In BCov two β-strands are paired only according to their alignment scores and the stability of conformations are not considered. Ruczinski et al. [7] showed that the arrangement of β-strands into β-sheets is not random. Based on the observations, there is a distinct pattern for β-strands arrangements. Some of the arrangements are unstable. Thus, they are never seen in

nature. On the other hand, some particular orientations are more favorable than others. In addition, models for computing the probability of open β-topologies for proteins were derived. The discriminative power of these models is reduced significantly because the number of possible β-strand organizations increase exponentially and there is not sufficient training data to reliably represent such conformations. Therefore, these models are limited to proteins that contain at most ten β-strands. In this research, we try to improve BCov by considering the stability and frequency of β-strand pairing and β-sheet conformation.

3. Proposed Method

In this article, two efforts are made to resolve the problem of predicting β-sheet topology: *BetaProbe1* and *BetaProbe2*.These efforts can predict both β-sheet topology and β-contact map. As previously mentioned, in BCov[26] two β-strands are paired based on only their alignment score; but, Ruczinski et al. [7] showed that the organization of β-strands into β-sheets is not random and there is a distinct pattern. Therefore, to improve BCov, we attempt to give greater chances to more stable and more frequent conformations during the selection. In this section, first, a general description of each attempt is presented. Then, the steps of the proposed methods are described in detail.

3-1. First Effort: BetaProbe1

BetaProbe1 consists of three major steps: (i) in order to achieve more accurate alignments, a dynamic programming approach is used to compute the β-strand pairwise alignment probability. In addition, pairwise interaction probability of each pair of β-strands is computed according to [32]. Then, both pairwise alignment probability and pairwise interaction probability are utilized to compute the score of each interaction (ii) to determine the maximum total strand-to-strand contact potentials of the protein an integer programming optimization is used. In this step, to enforce more stable and more observed paired β-strands to be selected, pairwise interaction scores obtained in the previous step are utilized (iii) the best β-sheet topology is achieved according to paired strands determined in the previous step. To predict more stable conformations, β-sheet topology probabilities are considered. The pseudo code of *BetaProbe1* is illustrated in Pseudo code1.

Computing β-strand Pairwise Interaction Score: Many methods have been proposed to find the best alignment between sequences [33], [34]. Here we concentrate on an alignment method which is especially proposed for β-strands. In *BetaProbe1* the alignment probability of each two β-strands is computed based on the proposed method in BetaZa[25]. In this method, the Needleman-Wunsch algorithm [33][34] is used to compute the optimum alignment between each pair of β-strands in the parallel and anti-parallel directions. Then, the probability of the optimum alignment is computed by dividing the score of the best alignment by the sum of all possible alignments. To improve the accuracy of the alignments, the amino acid pairing potentials are used which are computed especially based on the β-amino acids.

Pseudocode 1: Probability-based algorithm for β-sheet topology prediction (*BetaProbe1*)

- ***Input:*** *protein's strands*
- ***Output:*** *an open β-sheet conformation with the highest probability*

- ***Step 1:*** *Determining β-strand Pairwise Interaction Score*
  - ***for*** *each pair of strands s_i and s_j* ***do***
    - ***compute*** *their parallel and anti-parallel pairwise alignment probabilities*
    - ***compute*** *their parallel and anti-parallel pairwise interaction probabilities*
    - ***scores****=alignment probability × interaction probability*
- ***Step 2:*** *Predicting the Closed β-Sheet Topology*
  - ***Solve*** *the integer programming problem*
- ***Step 3:*** *Determining the Best Open β-Sheet Topology*
  - ***for*** *each closed β-topology* ***do***
    - ***for*** *each interaction between two β-strands* ***do***
      - ***Omit*** *the interaction temporarily*
      - ***Compute*** *the probability of the new open β-sheet*
    - ***Select*** *the open β-sheet with the highest conformation probability.*

To store the pairwise alignment probability, a matrix called "PAP (Pairwise Alignment Probability)" with n rows and 2n columns is defined. In this matrix, n is the number of β-strands in the protein. Matrix PAP is defined as follows:

$$PAP(i,j)=\begin{cases} S_{parallel}(s_i,s_j) & \text{if } i\le n \text{ and } j\le n \text{ and } j\neq i \\ S_{anti\text{-}parallel}(s_i,s_j) & \text{if } i\le n,\ n+1\le j\le 2\times n \text{ and } j\neq n+i \\ 0 & \text{if } j=i \text{ or } j=n+i \end{cases} \quad (1)$$

In Equation (1), $S_{parallel}$ (s_i,s_j) represents the probability of optimum alignment between strands s_i, i=1,2,…,n, and s_j, j=1,2,…,n, where their interaction type is parallel. Also, $S_{anti\text{-}parallel}$ (s_i,s_j) represents the probability of optimum alignment between strands s_i, i=1,2,…,n, and s_j, j=1,2,…,n, where their interaction type is anti-parallel. The definition shows that the matrix PAP is divided into two sections with an equal number of columns. The left section is used to store the parallel alignment probabilities and the right section is used to store the anti-parallel ones. The Score matrix for the protein in Fig. 1 is shown in Fig. 2-(a). It is important to note that the alignment probability depends on the spatial ordering of strands [25]. Therefore, the score of the optimum alignment between non-bridge strands can be different. This is expressed in (2) and (3):

$$S_{parallel}(s_i,s_j)\neq S_{parallel}(s_j,s_i) \quad (2)$$

$$S_{antiparallel}(s_i,s_j) \neq S_{antiparallel}(s_j,s_i) \quad (3)$$

According to [32], some β-strand pairs are more stable and they are more frequently observed in nature, as compared to others. Based on this observation, matrix "PIP (Pairwise Interaction Probability)" is defined to store the pairwise interaction probabilities of β-strands. The models derived by [32] were used to compute these probabilities. Matrix PIP contains n rows and 2×n columns as defined in (4):

$$PIP(i,j)=\begin{cases} P_{parallel}(s_i,s_j) & \text{if } i\leq n \text{ and } j\leq n \text{ and } j\neq i \\ P_{antiparallel}(s_i,s_j) & \text{if } i\leq n,\ n+1\leq j\leq 2\times n \text{ and } j\neq i+n \\ 0 & \text{if } j=i \text{ or } j=i+n \end{cases} \quad (4)$$

In (4), $P_{parallel}(s_i,s_j)$ represents the probability of strands s_i and s_j to make a parallel interaction in the final structure based on the protein characteristics such as the helical status and the number of residues between each two beta strands. Similarly, $P_{antiparallel}(s_i,s_j)$ is the probability of strands s_i and s_j to make an antiparallel interaction. The spatial ordering of strands has no effect on the β-strand pairwise interaction probability. This is expressed in (5) and (6):

$$P_{parallel}(s_i,s_j) = P_{parallel}(s_j,s_i) \quad (5)$$

$$P_{antiparallel}(s_i,s_j) = P_{antiparallel}(s_j,s_i) \quad (6)$$

In Fig. 2-(b), the matrix PIP is computed for the protein 1NZ0D. Then the scores of interactions between each pair of β-strands are determined. In the computation of each score, both pairwise interaction probability and pairwise alignment probability is considered. To store the scores of the interactions, a bi-dimensional n×2n matrix called "Score" is introduced, where n is the number of β-strands in the protein. The matrix definition is declared in (7).

$$Score(i,j)=PAP(i,j)\times PIP(i,j)\ 1\leq i\leq n\ ,\ 1\leq j\leq 2\times n \quad (7)$$

In the matrix Score definition, the first n columns represent the scores of parallel interactions. Similarly, the last n columns show anti-parallel ones. It is important to note that the score of an interaction between two strands depends on their spatial ordering. The Score matrix is illustrated in Fig.2-(c) for the protein 1NZ0D.

Prediction of the Closed β-Sheet Topology: Unlike BCov, in the integer optimization problem the pairwise interaction probabilities are considered in order to predict more stable paired β-strands. In addition, the integer programming model of *BetaProbe1* is defined differently from the BCov's. As a result, the closed β-sheet topology is obtained by solving the integer problem in (8).

$$\begin{aligned} &\textit{maximize:} && \sum_{i=1}^{n}\sum_{j=1}^{2\times n} Score(i,j)\ X(i,j) \\ &\textit{subject to:} && c1:\ X(i,j)\in\{0,1\}\ \forall\ 1\leq i\leq n\ ,\ 1\leq j\leq 2\times n \\ & && c2:\ X(i,j)+X(j,i)+X(i,j+n)+X(j,i+n)\in\{0,1\} \\ & && \qquad \forall\ 1\leq i\leq n,\ 1\leq j\leq 2\times n \\ & && c3:\ \textstyle\sum_{j=1}^{2\times n} X(i,j)\in\{0,1\}\ \forall 1\leq i\leq n \\ & && c4:\ \textstyle\sum_{i=1}^{n}\big(X(i,j)+X(i,j+n)\big)\in\{0,1\}\ \forall 1\leq j\leq n \\ & && c5:\ \textstyle\sum_{j=1}^{2\times n} X(i,j)+\sum_{j=1}^{n}(X(j,i)+X(j,i+n))\in\{1,2\} \\ & && \qquad \forall 1\leq i\leq n \\ & && c6:\ X(i,i)=X(i,i+n)=0\ \forall 1\leq i\leq n \end{aligned} \quad (8)$$

(a)

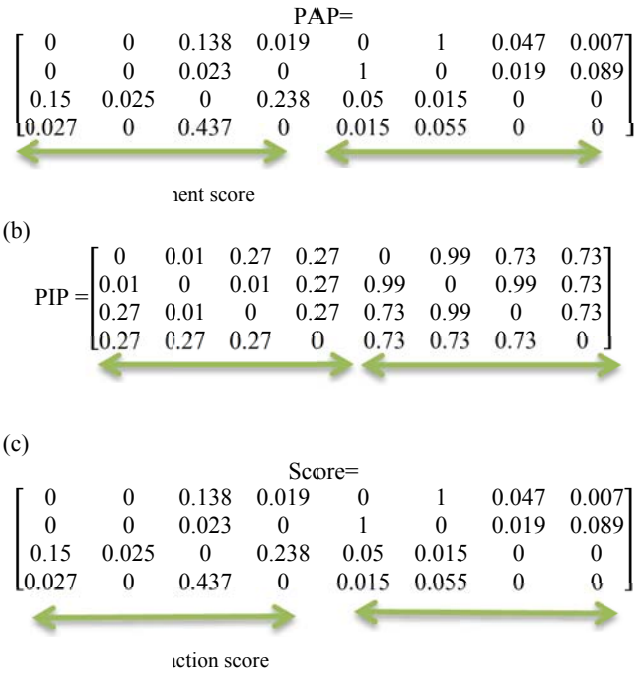
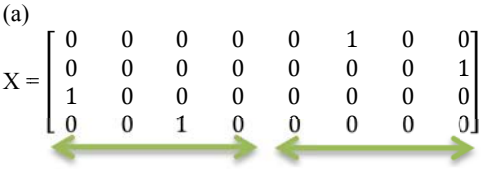


Fig 2. (a) The matrix PAP for a protein with PDB ID 1NZ0D computed by the dynamic programming algorithm [25]. (b) The matrix PIP for protein 1NZ0D computed by using the pairwise interaction probabilities in [32]. (c) The matrix Score for protein 1NZ0D computed by considering both pairwise alignment probability and pairwise interaction probability in this paper.

X is a n×2n binary matrix (constraint c1) in which non-zero entries show an interaction between two related strands. c2 constraint shows whether the interaction between two strands is parallel or antiparallel. c3 and c4 constraints ensure that all strands have at most one strand partner on either side. Furthermore, each strand can pair with at least one and at most two other β-strands (constraint c5).

In Fig. 3, the matrix X and the predicted closed β-sheet topology for protein 1NZ0D are shown.

(a)



(b)

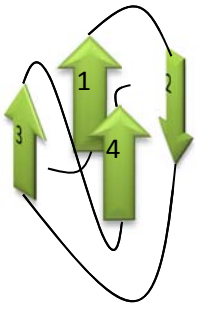


Fig. 3. (a) The matrix X obtained by solving the integer program. (b) The predicted closed β-sheet topology for the protein in Fig. 1.

Determining the Best Open β-Sheet Topology: In the previous step, paired β-strands and their interaction types are determined by the integer program solution. The predicted interactions make closed β-sheets, in other words, each strand has two partners. To extend the proposed method for the open β-sheets, the β-sheet topology probabilities determined by [32] are used. In this step, the probability of each possible open sheet is computed. Then the most probable one is selected as the best β-sheet topology. To enumerate all possible open β-sheets, one of the interactions of the closed one is omitted at a time. The process of determining the best open β-sheet topology is illustrated in Fig.4. In addition, Fig. 5 shows all possible open β-sheets for the closed one in the Fig. 3-(b).

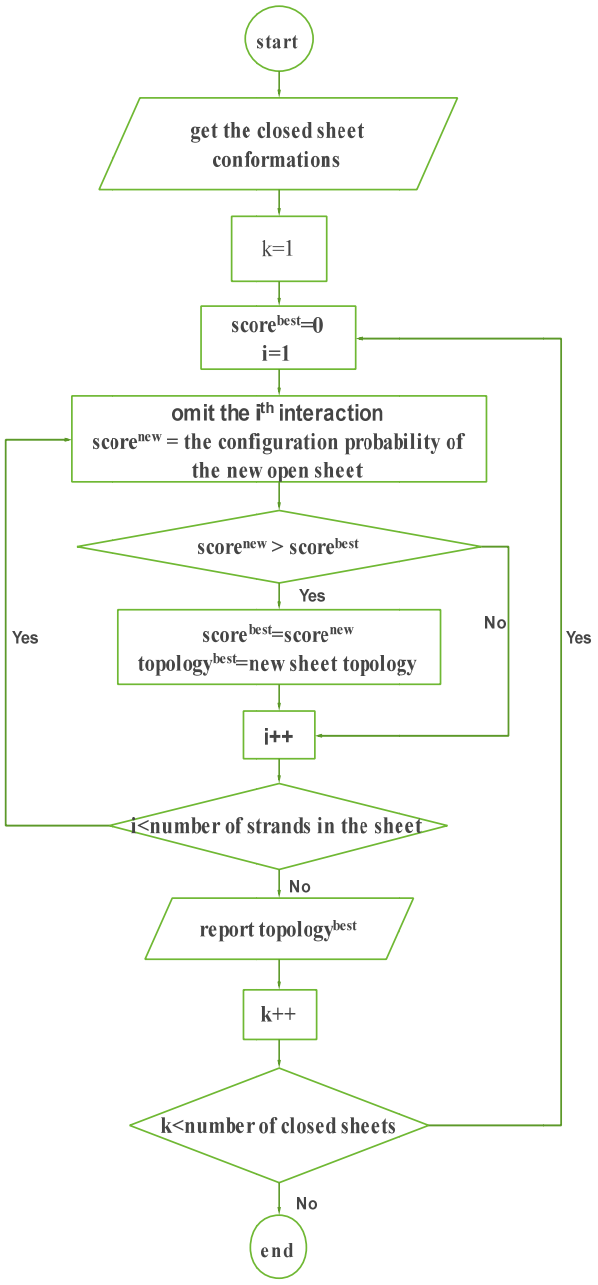


Fig. 4. The process of determining the most probable open β-sheet

3-2. Second Effort: BetaProbe2

BetaProbe2 consists of two major steps: (i) similar to *BetaProbe1*, the score of each interaction is computed by considering both pairwise alignment probability and pairwise interaction probability. To obtain more accurate alignments, a dynamic programming approach is used to compute the alignment probability of each pair of β-strands (ii) to unravel the problem of determining the β-sheet topology, an integer programming optimization is introduced. Unlike *BetaProbe1*, the integer problem is defined to maximize the product of the interaction scores. In this step, both pairwise interaction probabilities and pairwise alignment probabilities are utilized to give a greater chance to more stable and more observed paired β-strands for selection. Unlike *BetaProbe1*, the β-sheet topology achieved by the integer program solution is not closed. The pseudo code of *BetaProbe2* is illustrated in Pseudocode 2.

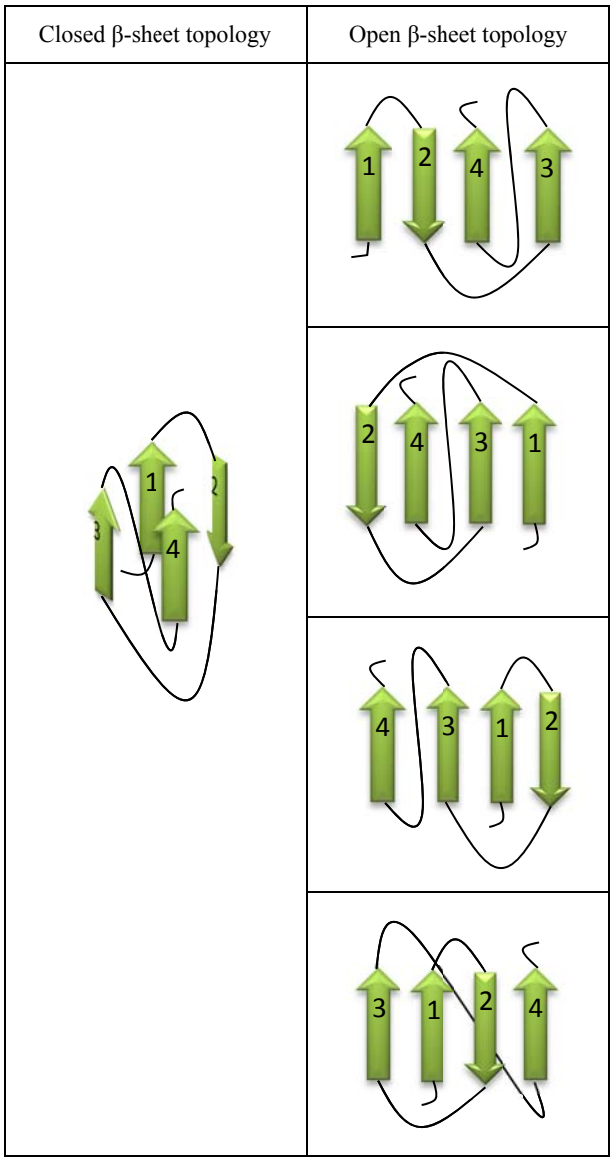


Fig. 5. All possible open β-sheets from a closed one. The gray cell shows the best open β-sheet topology for protein 1NZ0D.

Pseudocode 2: Probability-based algorithm for β-sheet topology prediction (*BetaProbe2)*

- ***Input:*** *protein's strands*
- ***Output:*** *an open β-sheet conformation with the highest probability*

- ***Step 1:*** *determining B-strand pairwise Interaction Score*

  for *each pair of strands* s_i *and* s_j ***do***

  compute *their parallel and anti-parallel pairwise alignment probabilities*

  compute *their parallel and antiparallel pairwise interaction probabilities*

  scores*=alignment probability×interaction probability*

- ***Step 2:*** *Prediction of the β-Sheet Topology*

  for *each pair of strands* s_i *and* s_j ***do***

  $scores^{new}=log(scores)$

  Solve *the integer programming problem*

Computing β-strand Pairwise Interaction Score: Similar to *BetaProbe1*, first the elements of matrices PIP and PAP are computed as in *BetaProbe1*. Then, the score of interaction between each pair of β-strands is determined.

Determining the β-sheet Topology: The problem of specifying the best β-sheet topology is reduced to an integer optimization. By assuming that the event of existing an interaction between two strands is independent of other β-strand interactions, the probability of the occurrence of several interactions is computed by the product of their probabilities. Therefore, an integer optimization is used to maximize the product of β-strand pairwise interaction scores, because each pairwise interaction score shows the probability of occurrence of an interaction between two strands according to the pairwise alignment probability and the pairwise interaction probability. Since pairwise interaction scores have positive values and the logarithm function is ascending, it is possible to maximize the sum of the logarithms of the pairwise interaction scores instead of maximizing their product. Then, the problem of determining the β-sheet topology becomes an integer linear problem represented in (9). Note that the constraints of the problem are the same as (8). In Fig. 6, the matrix X and the final β-sheet topology for protein 1NZ0D are presented.

$$\begin{aligned}
&\textit{maximize:} \quad \sum_{i=1}^{n}\sum_{j=1}^{2\times n} log(Score(i,j))\,X(i,j) \\
&\textit{subject to:} \quad c1:\ X(i,j)\in\{0,1\}\ \forall\, 1\le i\le n,\ 1\le j\le 2\times n \\
&\qquad c2:\ X(i,j)+X(j,i)+X(i,j+n)+X(j,i+n)\in\{0,1\} \\
&\qquad\qquad \forall\, 1\le i\le n,\ 1\le j\le 2\times n \\
&\qquad c3:\ \textstyle\sum_{j=1}^{2\times n} X(i,j)\in\{0,1\}\ \forall 1\le i\le n \\
&\qquad c4:\ \textstyle\sum_{i=1}^{n}\big(X(i,j)+X(i,j+n)\big)\in\{0,1\}\ \forall 1\le j\le n \\
&\qquad c5:\ \textstyle\sum_{j=1}^{2\times n} X(i,j)+\sum_{j=1}^{n}(X(j,i)+X(j,i+n))\in\{1,2\} \\
&\qquad\qquad \forall 1\le i\le n \\
&\qquad c6:\ X(i,i)=X(i,i+n)=0\ \forall\, 1\le i\le n
\end{aligned} \tag{9}$$

(a)

$$X=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

arallel

(b)

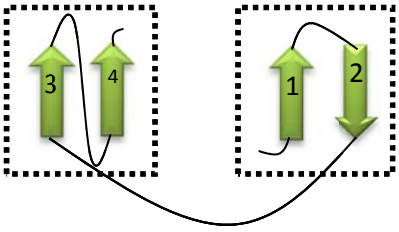


Fig 6. (a) matrix *X* represent the result of solving integer programming problem. (b) the final predicted β-sheet topology

4. Results

In this section, first, the evaluation metrics and the data set are described. Then, the results of evaluating *BetaProbe1* and *BetaProbe2* are presented.

Evaluation metrics: To evaluate the performance of the proposed methods, well-known metrics in (10), (11) and (12) are used. These metrics have been used to evaluate state-of-the-art methods [5], [26], [25]:

$$Precision=\frac{TP}{Tp+FP}\times 100 \tag{10}$$

$$Recall=\frac{TP}{TP+FN}\times 100 \tag{11}$$

$$F1\text{-}score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{12}$$

Note that TP, FP, and FN represent true positives, false positives, and false negatives values, respectively.

Dataset: We used the BetaSheet916 set for the evaluation. This dataset is extracted from the PDB by [4]. It includes 916 proteins. To perform cross-validation, it is split into 10-folds randomly and evenly. DSSP program [35] is used for assigning the secondary structure. In this article β-residues includes: (1) the extended β-strands (shown by E in the DSSP) and (2) the isolated β-bridges (shown B in the DSSP output).

Cross validation: At each step in a cross-validation, one fold is considered as the test data and the remaining ones are the training set. Models are trained based on the training set. Predictions are determined in the test set. This process is repeated for all proteins in the original set. The accuracy measures are computed after the predictions are accomplished.

We carried out three simulations. In the first simulation, a 10-fold cross-validation experiment was performed on the BetaSheet916 for proteins with less than or equal to four β-strands and less than three partners. Similarly, in the second simulation, the proposed method was evaluated on proteins with less than or equal to five β-strands with less than three

partners. The third simulation was performed on proteins with less than or equal to six β-strands with less than three partners.

BetaProbe1 and *BetaProbe2* were compared with the state-of-the-art method, BCov, which is also based on integer programming. For this purpose, in the first step of BCov, the residue pairing probabilities calculated in BetaPro were used. Then the methods were evaluated on the same data set.

In Table 1, the performance of *BetaProbe1* at the strand level is compared with the performance of BCov. The recall, precision, and F1-score measures are shown in this table.

Wilcoxon test for related samples has been utilized to determine whether there is a significant difference in the precision, recall, and F1-score of the two methods. To perform the test, the data set was broken into ten subsidiaries as declared in the Dataset section. After that, the results of *BetaProbe1* and BCov were evaluated for these subsets. The test showed that with an average error of 5%, there is a significant difference between the recall of the two methods at the pairing direction level for proteins with up to six and up to five strands. This means that the recall improvement of *BetaProbe1* compared with that of BCov is significantly meaningful. From Table 1, it can be concluded that besides using β-sheet conformation probabilities, considering pairwise interaction probabilities in the computation of β-strands interaction score and combining it with the integer programming greatly improves the accuracy of pairing directions. In Chart 1, Chart 2, and Chart 3 the recall, precision, and F1-score of *BetaProbe1* at pairing direction level is illustrated and compared to BCov's.

Table 1. The performance of *BetaProbe1* at strand level on proteins with 6 or fewer β-strands on BetaSheet916.

Evaluation level	Method	Recall	Precision	F1-score
strand pairing	BCov≤6 [a]	79	84	82
	BetaProbe1≤6	73	69	71
	BCov≤5 [b]	81	86	83
	BetaProbe1≤5	76	73	74
	BCov≤4 [c]	82	85	83
	BetaProbe1≤4	75	72	74
Pairing direction	BCov≤6	64	68	66
	BetaProbe1≤6	70	67	68
	BCov≤5	64	69	66
	BetaProbe1≤5	73	70	72
	BCov≤4	72	75	73
	BetaProbe1≤4	74	70	72

a)The evaluation is done on proteins with up to 6 β-strands
b) The evaluation is done on proteins with up to 5 β-strands
c) The evaluation is done on proteins with up to 4 β-strands

In Table 2, the performance of *BetaProbe2* is compared to BCov at strand level. For the three subsets of proteins, the precision and the F1-score measures of the proposed method at pairing direction level is better than BCov. Further, the precision of *BetaProbe2* is better than BCov at strand pairing level. Wilcoxon test for related samples showed that with an average error of 5%, there is a significant difference between the precision of BCov and *BetaProbe2* at the pairing direction level for all subsets of proteins. It can be concluded that the precision improvement of *BetaProbe2* is significantly meaningful as compared with BCov's.

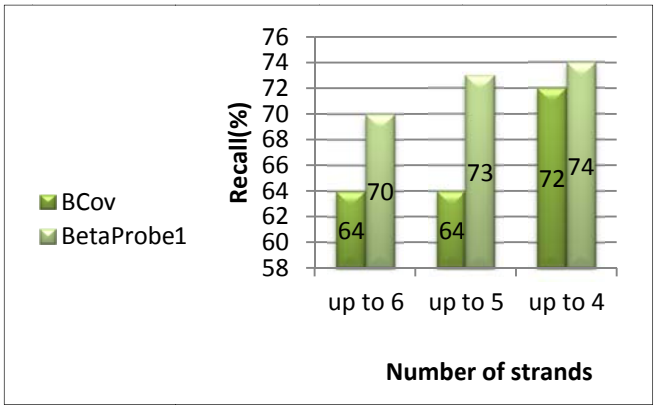


Chart 1: Recall comparison of *BetaProbe1* to BCov

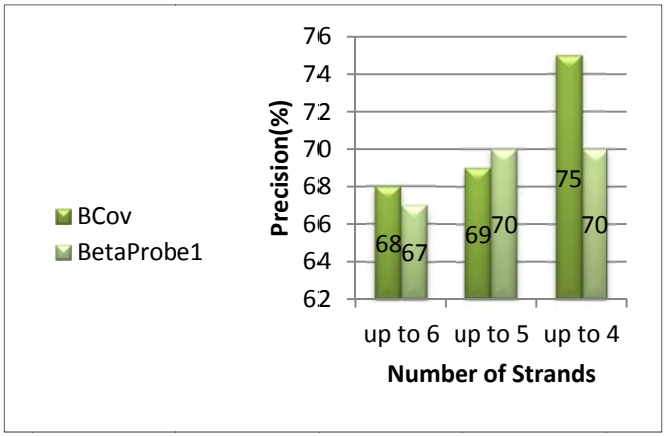


Chart 2: Precision comparison of *BetaProbe1* to BCov

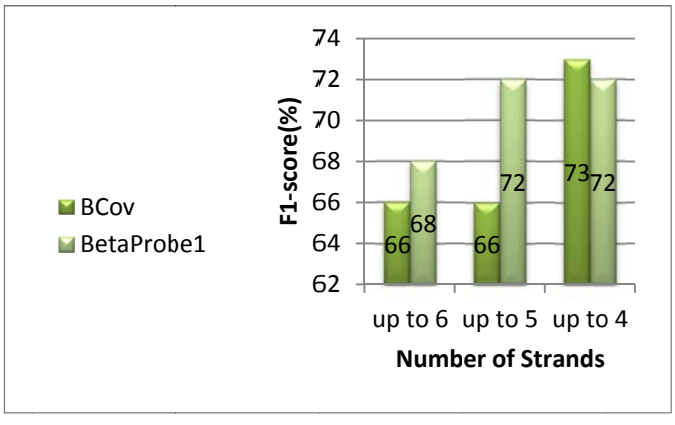


Chart 3: F1-score comparison of *BetaProbe1* to BCov

The reason for the improvement of proposed methods as compared with BCov is that adding pairwise interaction probabilities to the integer programming in the second step, enforces β-strand interactions which are more frequent in the nature to be selected with higher probabilities. In addition, to improve the pairwise alignments between β-strands, a dynamic programming approach is utilized in which gaps are allowed. Furthermore, using the amino acid pairing potentials provided by the BetaZa in the first step has improved the accuracy.

Table 2. The performance of *BetaProbe2* at strand level on proteins with 6 or fewer β-strands on Beta Sheet 916.

Evaluation level	Method	Recall	Precision	F1-score
strand pairing	BCov≤6	79	84	82
	BetaProbe2≤6	65	85	73
	BCov≤5	81	86	83
	BetaProbe2≤5	68	87	76
	BCov≤4	82	85	83
	BetaProbe2≤4	71	90	79
Pairing direction	BCov≤6	64	68	66
	BetaProbe2≤6	63	83	72
	BCov≤5	64	69	66
	BetaProbe2≤5	67	87	75
	BCov≤4	72	75	73
	BetaProbe2≤4	71	89	79

In Table3 and Table4 the results of *BetaProbe2* are compared to BetaZa at the residue level and strand level, respectively. The same alignment technique is used in both methods. BetaZa searches the entire search space to find the best β-sheet topology. Although the execution time of *BetaProbe2* is less than BetaZa, the precision of *BetaProbe2* is better at pairing direction level. In addition, the precision of *BetaProbe2* at strand pairing level and contact map level is comparable with BetaZa's. Comparing the recall, precision, and F1-score measures of *BetaProbe2* at pairing direction level with the other method, the results are represented in Chart4, Chart 5, and Chart 6, respectively.

In Table5 and Table6 the performance of *BetaProbe1* and *BetaProbe2* are represented at the residue level and strand level, respectively. As mentioned before, in *BetaProbe1*, the sum of the interaction scores is maximized in the integer programming step while in *BetaProbe2* the product of the interaction scores is maximized. It leads to predicting fewer interactions between β-strands in *BetaProbe2* because the scores of the interactions are in the range of zero and one. Subsequently, the predicted pairwise interactions are the most frequent ones. Therefore the precision of predicted interactions increases while the recall decreases.

Table 3. The performance of BetaProbe2 at strand level on proteins with 6 or fewer β-strands on Beta Sheet 916

Evaluation level	Method	Recall	Precision	F1-score
Contact map	BetaZa≤6	78	80	79
	BetaProbe2≤6	58	80	67
	BetaZa≤5	80	80	80
	BetaProbe2≤5	60	79	68
	BetaZa ≤4	82	82	82
	BetaProbe2≤4	61	80	69

Table 4. The performance of *BetaProbe2* at strand level on proteins with 6 or fewer β-strands on Beta Sheet 916.

Evaluation level	Method	Recall	Precision	F1-score
strand pairing	BetaZa≤6	83	84	84
	BetaProbe2≤6	65	85	73
	BetaZa≤5	87	88	87
	BetaProbe2≤5	68	87	76
	BetaZa ≤4	91	91	91
	BetaProbe2≤4	71	90	79
Pairing direction	BetaZa ≤6	80	81	81
	BetaProbe2≤6	63	83	72
	BetaZa ≤5	84	85	84
	BetaProbe2≤5	67	87	75
	BetaZa ≤4	88	88	88
	BetaProbe2≤4	71	89	79

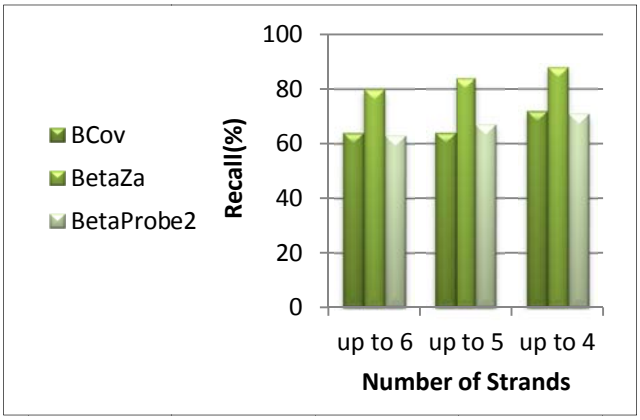


Chart 4: Recall comparison of *BetaProbe2* to other methods

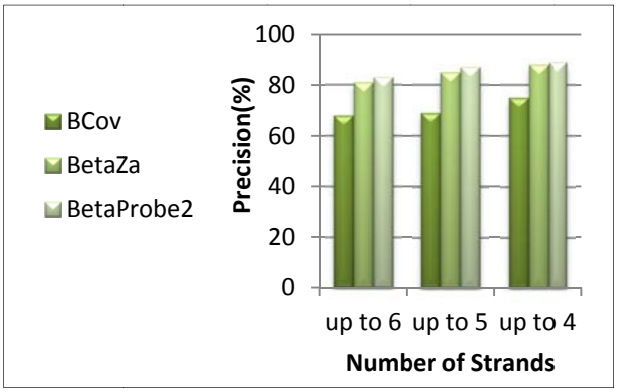


Chart 5: Precision comparison of *BetaProbe2* to other methods

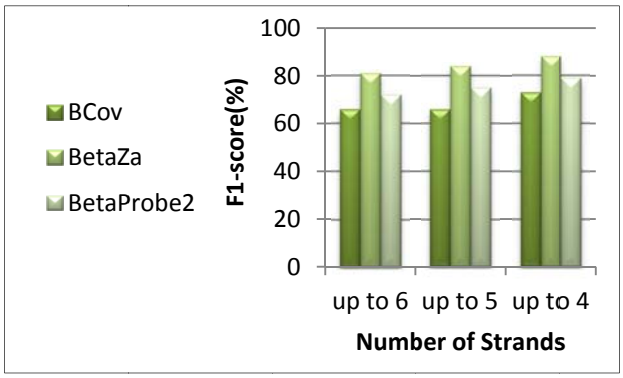


Chart 6: Precision comparison of *BetaProbe2* to other methods

5. Conclusion and Future Work

The issue of determining the topology of β-sheets is considered as a challenging problem. In this paper, *BetaProbe1* and *BetaProbe2*, two probability-based methods for the β-sheet topology prediction, are introduced. In these methods, first, the optimum pairwise alignment probabilities of β-strands are determined using the dynamic programming approach while any number of gaps are allowed. Then, the probability of the occurrence of an interaction is computed. After that, the score of on interaction is computed utilizing both pairwise alignment probability and pairwise interaction probability. Finally, we reduced the problem of finding the β-sheet topology to an integer optimization. 10-fold cross-validation experiments are performed to evaluate the proposed methods. The results show that these methods outperform the most recent integer programming-based method[26]. The major novelties in this research can be summarized as follow:

1. Considering both pairwise alignment probability and pairwise interaction probability to compute the score of an interaction between two β-strands;
2. Combining the probability of occurrence of an interaction with the integer programming;
3. Considering β-sheet conformation probability in the nature to predict more frequent β-topologies;
4. Considering the spatial ordering of β-strands in β-sheets in the integer programming;
5. The ability of the proposed methods to predict the β-sheet structure for proteins with multiple β-sheets;
6. The ability of the proposed methods to predict the β-sheet topology for proteins with closed β-sheets.

The performance of predictions can be improved even further. By combining residue pairing propensities with PSICOV [18] ones, the methods can become more accurate. Our methods can predict proteins with six or fewer β-strands with less than three partners. This can be extended to predict proteins with a higher number of β-strands and higher order partners by extending probabilities and adding new constraints to the integer programming step.

Table 5. The performance of *BetaProbe1* and *BetaProbe2* at residue level on proteins with 6 or fewer β-strands on Beta Sheet 916.

Evaluation level	Method	Recall	Precision	F1-score
Contact map	BetaProbe1≤6	63	66	64
	BetaProbe2≤6	58	80	67
	BetaProbe1≤5	64	66	65
	BetaProbe2≤5	60	79	68
	BetaProbe1≤4	64	67	65
	BetaProbe2≤4	61	80	69

Table 6. The performance of *BetaProbe1* and *BetaProbe 2* at strand level on proteins with 6 or fewer β-strands on Beta Sheet 916.

Evaluation level	Method	Recall	Precision	F1-score
strand pairing	BetaProbe1≤6	73	69	71
	BetaProbe2≤6	65	85	73
	BetaProbe1≤5	76	73	74
	BetaProbe2≤5	68	87	76
	BetaProbe1≤4	75	72	74
	BetaProbe2≤4	71	90	79
Pairing direction	BetaProbe1≤6	70	67	68
	BetaProbe2≤6	63	83	72
	BetaProbe1≤5	73	70	72
	BetaProbe2≤5	67	87	75
	BetaProbe1≤4	74	70	72
	BetaProbe2≤4	71	89	79

References

[1] J. Peng, "Statistical inference for template-based protein structure prediction," Doctoral thesis, Toyota Technological Institute at Chicago, 2013.

[2] C. W. O'Donnell, "Ensemble modeling of beta-sheet proteins," PhD thesis, Massachusetts Institute of Technology, 2011.

[3] M. J. Sternberg and J. M. Thornton, "On the conformation of proteins: an analysis of beta-

pleated sheets," *J. Mol. Biol.*, vol. 110, no. 2, pp. 285–296, 1977.

[4] J. Cheng and P. Baldi, "Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms," *Bioinformatics*, vol. 21, no. suppl 1, pp. i75–i84, 2005.

[5] A. Subramani and C. A. Floudas, "Beta-Sheet Topology Prediction With High Precision and Recall for Beta and Mixed alpha/beta Proteins," *PLoS One*, vol. 7, no. 3, 2012.

[6] S. M. Zaremba and L. M. Gregoret, "Context-dependence of Amino Acid Residue Pairing in Antiparallel β-Sheets," *J. Mol. Biol.*, vol. 291, no. 2, pp. 463–479, 1999.

[7] I. Ruczinski, C. Kooperberg, R. Bonneau, and D. Baker, "Distributions of beta sheets in proteins with application to structure prediction," *Proteins Struct. Funct. Bioinforma.*, vol. 48, no. 1, pp. 85–97, 2002.

[8] T. Kortemme, "Design of a 20-Amino Acid, Three-Stranded -Sheet Protein," *Science*, vol. 281, no. 5374, pp. 253–256, Jul. 1998.

[9] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, pp. 1364–1368, 2003.

[10] J. S. Merkel and L. Regan, "Modulating protein folding rates in vivo and in vitro by side-chain interactions between the parallel β strands of green fluorescent protein," *J. Biol. Chem.*, vol. 275, no. 38, pp. 29200–29206, 2000.

[11] Y. Mandel-Gutfreund, S. M. Zaremba, and L. M. Gregoret, "Contributions of residue pairing to β-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel β-strands," *J. Mol. Biol.*, vol. 305, no. 5, pp. 1145–1159, 2001.

[12] M. Eghdami, T. Dehghani, and M. Naghibzadeh, "BetaProbe: A probability based method for predicting beta sheet topology using integer programming," in *Computer and Knowledge Engineering (ICCKE), 2015 5th International Conference on*, 2015, pp. 152–157.

[13] A. N. Tegge, Z. Wang, J. Eickholt, and J. Cheng, "NNcon: improved protein contact map prediction using 2D-recursive neural networks," *Nucleic Acids Res.*, vol. 37, no. suppl 2, pp. W515–W518, 2009.

[14] J. Eickholt and J. Cheng, "Predicting protein residue–residue contacts using deep networks and boosting," *Bioinformatics*, vol. 28, no. 23, pp. 3066–3072, 2012.

[15] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, no. 1, pp. 113–121, 2007.

[16] D. Baú, A. J. M. Martin, C. Mooney, A. Vullo, I. Walsh, and G. Pollastri, "Distill: a suite of web servers for the prediction of one-, two-and three-dimensional structural features of proteins," *BMC Bioinformatics*, vol. 7, no. 1, p. 402, 2006.

[17] P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, 2012.

[18] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2012.

[19] Z. Wang and J. Xu, "Predicting protein contact map using evolutionary and physical constraints by integer programming," *Bioinformatics*, vol. 29, no. 13, pp. i266–i273, 2013.

[20] A. Kumar and L. Cowen, "Recognition of beta-structural motifs using hidden Markov models trained with simulated evolution," *Bioinformatics*, vol. 26, no. 12, pp. i287–i293, 2010.

[21] N. M. Daniels, R. Hosur, B. Berger, and L. J. Cowen, "SMURFLite: combining simplified Markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone," *Bioinformatics*, vol. 28, no. 9, pp. 1216–1222, 2012.

[22] N. M. Daniels, A. Gallant, N. Ramsey, and L. J. Cowen, "MRFy: remote homology detection for beta-structural proteins using Markov random fields and stochastic search," *Comput. Biol. Bioinformatics, IEEE/ACM Trans.*, vol. 12, no. 1, pp. 4–16, 2015.

[23] T. J. P. Hubbard, "Use of beta-strand interaction pseudo-potentials in protein structure prediction and modelling," in *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, 1994, vol. 5, pp. 336–344.

[24] R. E. Steward and J. M. Thornton, "Prediction of strand pairing in antiparallel and parallel beta sheets using information theory," *Proteins Struct. Funct. Bioinforma.*, vol. 48, no. 2, pp. 178–191, 2002.

[25] Z. Aydin, Y. Altunbasak, and H. Erdogan, "Bayesian models and algorithms for protein β-sheet prediction," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 8, no. 2, pp. 395–409, 2011.

[26] C. Savojardo, P. Fariselli, P. L. Martelli, and R. Casadio, "BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming," *Bioinformatics*, pp. 3151–3157, 2013.

[27] J. Jeong, P. Berman, and T. M. Przytycka, "Improving strand pairing prediction through exploring folding cooperativity," *IEEE/ACM Trans.*

Comput. Biol. Bioinforma., vol. 5, no. 4, pp. 484–491, 2008.

[28] M. Lippi and P. Frasconi, "Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights," *Bioinformatics*, vol. 25, no. 18, pp. 2326–2333, 2009.

[29] R. Fonseca, G. Helles, and P. Winter, "Ranking Beta Sheet Topologies with Applications to Protein Structure Prediction," *J. Math. Model. Algorithms*, vol. 10, no. 4, pp. 357–369, 2011.

[30] R. Rajgaria, Y. Wei, and C. A. Floudas, "Contact prediction for beta and alpha□beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD," *Proteins Struct. Funct. Bioinforma.*, vol. 78, no. 8, pp. 1825–1846, 2010.

[31] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[32] I. Ruczinski, "Logic regression and statistical issues related to the protein folding problem," PhD thesis, University of Washington, 2000.

[33] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[34] O. Gotoh, "An improved algorithm for matching biological sequences," *J. Mol. Biol.*, vol. 162, no. 3, pp. 705–708, 1982.

[35] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen□bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.