Journal of Computer Science

Review

Protein Sequences Features Extraction for Predicting Beta-Turns and their Types: A Review

[1,2]Murtada Khalafallah Elbashir Elfaki

[1]Department of Information Systems,
College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf, Kingdom of Saudi Arabia
[2]Department of Computer Sciences, Faculty of Mathematical and Computer Sciences, University of Gezira, Madani, Sudan

Article history
Received: 26-07-2019
Revised: 24-08-2019
Accepted: 30-09-2019

Email: murtadabashir@gmail.com

Abstract: Beta-turns are considered to be important secondary structure types that have essential role in molecular recognition, protein folding and stability. They represent 25% of protein structures, therefore they are the most common type of non-repetitive or tight turns structures. Their prediction is considered to be an important issue in bioinformatics and molecular biology, because it provides valuable information and inputs for the fold recognition and drug design. There are many machine learning and statistical based approaches that were designed to predict beta-turns. Among the successful approaches that are based on machine learning are the approaches that used Neural Networks (NNs) and Support Vector Machines (SVMs) methods. These approaches used different features and features organizations. Among the most usable features in beta-turns prediction are the Position Specific Scoring Matrices (PSSMs) and the predicted secondary structure. This work gives a review of the most successful methods that are used for beta-turns prediction and the features as well as the organizations of these features that they used.

Keywords: Beta-Turns, Position Specific Scoring Matrices, Predicted Secondary Structure, Predicted Shape String

Introduction

Secondary structure of proteins is considered to be an important topic in bioinformatics and it consists of alpha-helices, beta-sheets, random coils and turns. Alpha-helices and beta-sheets are considered as regular secondary structure, because they are sequences of residues with repeating φ and ψ values. The residues that correspond to turns structures do not form a regular secondary structure. In turns structures the Calpha-atoms of two residues are separated by one to five peptide bonds and the distance between these Calpha-atoms is less than 7A°. The number of peptide bonds that separate the two end residues determines the specific turn type. In alpha-turns and beta-turns, the two end residues are separated by four and three peptide bonds respectively. In gamma-turns, delta-turns and pi-turns, the two end residues are separated by two, one and five peptide bonds respectively.

Beta-turns are the most common type of turn structures since they represent 25% of the secondary structure of the protein sequence. They have the ability to bring together and allow the interaction between the regular secondary structures elements thus their prediction is of significance to protein folding (Petersen *et al*., 2010). Beta-turns are also important in the biological activities of peptides as the bioactive structures that interact with other molecules such as receptors, enzymes and antibodies and they are important in the design of various peptidomimetics for many diseases (Kee and Jois, 2003; Zheng and Kurgan, 2008). Therefore, the prediction of beta-turns is important for providing valuable insights and inputs for the fold recognition as well as drug design. The beta-turns are not only two states classification problem but it can be further classified to 9 types according to the dihedral angles of residues $i + 1$ and $i + 2$ in the turn structure (Hutchinson and Thornton, 1994). The following Table 1 shows the dihedral angles of beta-turns types.

DOI: 10.3844/jcssp.2019.1331.1340

Table 1: The average values of the dihedral angles of β-turn types

	Dihedral angles (°)			
Turn type	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
I	-60	-30	-90	0
I'	60	30	90	0
II	-60	120	80	0
II'	60	-120	-80	0
IV	-61	10	-53	17
VIa1	-60	120	-90	0
VIa2	-120	120	-60	0
VIb	-135	135	-175	160
VIII	-60	-30	-120	120

The methods that are used for beta-turns prediction can be categorized as statistical based methods and machine learning based methods. The statistical based methods include (Chou and Fasman, 1974; Wilmot and Thornton, 1988; 1990; Chou, 1997; Chou and Blinn, 1997; Zhang and Chou, 1997; Fuchs and Alix, 2005). The machine learning methods are found to be the most successful methods, because they can handle the nonlinearity in the data very well. Most of the successful machine learning methods that are used for beta-turns prediction are based on Neural Networks (NNs), Support Vector Machines (SVMs) and k-nearest neighbour methods. The methods that use NNs include McGregor *et al.* (1989), BTPRED (Shepherd *et al*., 1999), BetaTpred2 (Kaur and Raghava, 2003), MOLEBRNN (Kirschner and Frishman, 2008) and NetTurnP (Petersen *et al*., 2010) and that which use SVMs methods include BTSVM method (Pham *et al*., 2003), the work of Zhang *et al*. (2005), Zheng and Kurgan's (2008), Hu and Li's (2008), the method of Liu *et al*. (2009), DEBT (Kountouris and Hirst, 2010), the method of Tang *et al.* (2011), our own work H-SVM-LR (Elbashir *et al*., 2013a) and Nguyen *et al*. (2014). The methods that are based on k-nearest neighbour include the work of Kim's (2004).

The features are very important inputs for prediction or classification using machine learning or statistical methods. Extracting or selecting the most informative features leads to high classification performance. Selecting the most informative features requires the experimentation of many features. Also some of the features may be combined together to enhance the accuracy of the machine learning methods. As shown in the previous paragraph, there are many researches that developed methods or techniques for beta-turns prediction. These methods used different features and features combinations. The common used features is the Position Specific Scoring Matrices (PSSMs). Since there is intercorrelation between various structural features of protein, secondary structure information has been widely used as an additional features and it enhances the prediction accuracy substantially. Recent researches added other features such as surface accessibility, predicted protein block, predicted backbone dihedral angle and predicted shape string.

Dataset and Performance Measures

There are many datasets that are used for the evaluation of beta-turns prediction methods. The most commonly used dataset in almost all of the recent researches is BT426 dataset therefore, the results that are pointed out in this paper are based on it. BT426 dataset has 426 non-homologous protein chains. It was developed by Guruprasad and Rajkumar (2000). X-ray of crystallography at two resolution or better is used to determine the structures of all the proteins chains in BT426 dataset. Each of these chains contains at least one beta-turns structure. 24.9% (approximately 25%) of all amino acids in BT426 have a beta-turns structure. The dataset can be downloaded from the link http://crdd.osdd.net/raghava/bteval/. The most frequently measures that are used to evaluate beta- turns prediction methods are the prediction accuracy and Matthew's correlation coefficient (MCC). It is important to use MCC with the accuracy because of the imbalanced dataset (25% beta-turns versus 75% non-beta-turns), where it is possible to achieve an accuracy of 75% by predicted all the residues to be non-beta-turns. In this paper the results of the prediction methods are based on these two measures.

Basic Sequence Information and Reliability Indices

The basic sequence information are normally obtained by encoding the protein sequence such that every amino acid type is represented by a single one according to its position in a row composed of 20 positions that represents the 20 amino acids, e.g., alanine, which is located in the first position is represented by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0. The predicted three secondary structures alpha helix, beta-sheets and coil are normally encoded as 1,0,0 and 0,1,0 and 0,0,1 respectively. The basic sequence information in addition with a secondary structure information which obtained using the reliability indices is used by Shepherd *et al*. (1999) to predict beta-turns. The reliability indices are the strengths of the prediction for each of the three target secondary structures states. They are represented by three integers in the range 0-9 and each index is divided by 10 to get three real numbers between 0 and 1 for each predicted secondary structure state. Shepherd *et al*. used a window size of 9 with 23 network inputs per window position (20 for the amino-acid information plus 3 for the secondary-structure information) to accomplish the prediction task. With filtering strategy, the accuracy and MCC of their methods are 74.9, 0.348 respectively. Shepherd *et al*. used a window size of

four on amino acid information only as an input to single layer net to predict the different types of beta-turns, they obtained an accuracy and MCC pairs of (91.2, 0.219), (95.5, 0.253), (95.7, 0.062), (96.8, 0.033) on types I, II, VIII, IV respectively.

PSSMs and Predicted Secondary Structure

PSSMs are M by 20 matrices, where M represents the sequence length and the number 20 refers to the 20 amino acids positions. These matrices are normally generated using PSI-BLAST (Altschul *et al*., 1997) using many rounds against specific sequence database. The widely used sequence database for generating them for the purpose of beta-turns prediction is the (NCBI) none redundant (nr) database. Beta-turns prediction is enhanced significantly using PSSM therefore many researchers used it in their prediction methods whether alone or in combination with other features. Kirschner and Frishman (2008) used two neural networks that derived from the Elman network (Elman 1990). These two NNS are used in ensemble manner in which the first NN is fed with PSSM of the sequence. The output of the first layer is fed to the second layer (structure to structure) to recognize beta-turns. They utilized a post-scaling by applying adjustable threshold on the output of the network to filter the prediction. The post-scaling (Montavon *et al*., 2012) is used to handle the unbalanced class distribution problem. The two stages neural network that is used by Andreas Kirschner and Dmitrij Frishman was adopted in bioinformatics by (Qian and Sejnowski, 1988; Rost and Sander, 1993; Adamczak *et al*., 2005). Their two stages neural networks obtained accuracy and MCC of 77.9 and 0.45 respectively. For the beta-turns types they used the threshold to obtain two results for their prediction one that maximizes the MCC and the other is a tradeoff between the MCC and the accuracy and their results are as follows: For type I by maximizing the MCC, the accuracy = 82.5 and MCC = 0.317, by using the tradeoff, the accuracy = 85.4 and MCC = 0.314. For type II the results of maximizing the MCC and the trade of are the same, the accuracy = 96.2 and MCC = 0.339. For type VIII by maximizing the MCC, the accuracy = 53.4 and MCC = 0.109, by using the tradeoff, the accuracy = 93.0 and MCC = 0.076. For type IV by maximizing the MCC, the accuracy = 72.3 and MCC = 0.236, by using the tradeoff, the accuracy = 85.2 and MCC = 0.196. For type I', the results of maximizing the MCC and the trade of are the same, the accuracy = 98.8 and MCC = 0.356. And for type II', the results of maximizing the MCC and the trade of are the same as well, the accuracy = 98.6 and MCC = 0.137.

Almost all of the recent researches combined the predicted secondary structure with PSSM to enhance the prediction accuracy. These researches used different secondary structure organizations. Zhang *et al*. (2005) used a window size of 7 on the PSSM and then added the secondary structure prediction. The total features that they used is 143. The accuracy and MCC that they obtained are 74.8% and 41% respectively.

Kaur and Raghava (2003) used two feed-forward back- propagation networks with a single hidden layer. They used a window size of 9 residue on the PSSM as an input to their networks. Both of their networks have a single hidden layer with 10 units. The prediction results of their first network, which is turn/non-turn (0 or 1) is combined with the probabilities of the predicted secondary structure (3 states) to form 4 units code, which is used as an input for the second network. The three structures states are provided by PSIPRED prediction method (Jones, 1999). The probabilities represent the strength of the prediction and they are in the range of 0-1. Harpreet Kaur and Gajendra Pal Singh Raghava filtered the final result of the prediction using a filtering strategy to calculate the final accuracy and MCC, which are found to be 75.5% and 43% respectively.

Zheng and Kurgan (2008) developed a method, which is considered to be the first to break the 80% accuracy barrier. They used sliding window of 7 to extract the features from PSSM. They employed four prediction methods to obtain the secondary structure features. These four methods are PSIPRED v2.5 (McGuffin *et al*., 2000; Bryson *et al*., 2005), JNET (Cuff and Barton, 2000), TRANSSEC (Montgomerie *et al*., 2006) and PROTEUS2 (Montgomerie *et al*., 2006). Each one of the four methods produces 3 features, where each represent specific structural states. The total number of the features generated from the four methods is 3×4 = 12. The confidence score of each one of the four prediction methods is added to the features vector after dividing it by 10. The confidence score added another 4 features to the feature vector. A binary value representing a specific arrangement of the secondary structure predicted with the four prediction methods for the central and the two adjacent residues is considered in the features vector. This binary value is calculated as follows: If the central amino acids is predicted as C then the two adjacent residues can be C and C this will form CCC arrangement, if one of them is C and the other one is either H or E and if X is assumed to be the set (E, H) then the resulted arrangement will be CCX, or XCC. If both of the adjacent residues are not C, this will result in the arrangement XCX. The total number of features produced by the binary number which represents specific arrangement of the secondary structure will be 4(number of prediction methos) * 3 (the three secondary structures states) * 4 (the pattens CCC, CCX, XCC, or XCX), which is equals to 48 features. Lastly Zheng C, Kurgan added the ratio between the number of residues in a given secondary structures and the window size, this will add additional 12 features. The same features organization that is used by Zheng C, Kurgan L is adopted by Elbashir *et al*. (2013b) for

predicting beta-turns in protein using Kernel logistic regression. Elbashir *et al*. obtained accuracy and MCC of 80.7 and 0.50 respectively.

Our own method (Elbashir *et al*., 2013a) used PSSMs and predicted secondary structure to predict the beta-turns. Because the training sets used for beta-turns prediction are imbalanced sets 1:3 for beta-turns and non-beta-turns, a clustered model is used. In the cluster model the non-beta-turns are clustered into 3 clusters and each cluster is used with the beta-turns cluster to form a balanced set that can be used to train three localized SVMs. Each localized SVM produce beta-turns and non-beta-turns predictions. The outputs of the three SVMs are combined to form a single beta-turn/non-turn output using fractional polynomial. The method tried different PSSMs and secondary structure organization i.e., using a sliding window on the PSSMs and then add the predicted secondary structure or using a sliding window on both PSSMs and predicted secondary structure. It was found that using sliding window on both PSSMs and predicted secondary structure produces the best results. Our own method obtained an accuracy and MCC of 82.87 and 0.56 respectively.

PSSMs, Predicted Secondary Structure and Surface Accessibility

Petersen *et al*. (2010) designed a method that consists of two layers of artificial neural networks. They utilized PSSMs, predicted secondary structure and surface accessibility, which is the surface area of a biomolecule that is accessible to a solvent as an input for the first layer Networks. The first layer networks consists of five network one of them is to predict whether an amino acid has beta-turn confirmation or not and the other four are used for predicting the position of the amino acid in the beta-turns confirmation (position1, position2, position3, position4). The surface accessibility is predicted using NetSurfP (Petersen *et al*., 2009), NetsurfP uses primary network that accepts PSSMs and secondary structure and produces 'B/E Classification' which refer to the raw buried/exposed. The output of the primary network is used with the PSSMs to form an input for the secondary network, which predict the buried/exposed of the given amino acid. The output from the first layers networks, which compose of five networks is used again with the secondary structure and surface accesability as an input to the second layer network to produce the final beta-turn/non-turn prediction. The method of Petersen *et al*. reached accuracy and MCC of 78.2% and 50% respectively. Petersen *et al*. used the same method of the first NN layer to predict the beta-turns types. Their prediction shows a MCC of 0.36, 0.23, 0.31, 0.16, 0.27, 0.16 on the types I, I', II, II', IV and VIII respectively.

PSSMs, Predicted Backbone Dihedral Angle and Secondary Structure

There is a high correlation between backbone dihedral angles and the secondary structure elements of the protein so they can be combined together in a feature matrix to enhance the predictions. Kountouris and Hirst (2010) added another features to the PSSM and predicted secondary structure, which is the seven state predicted dihedral angle that obtained from DISSPred (Kountouris and Hirst, 2009). DISSPred is also used to predict the three state secondary structures elements. They used a sliding window of nine on the PSSM to obtain (9×20 = 180) dimension vector. The window size that is used on the three predicted secondary structures states and the seven state predicted dihedral angle is five. These features add (3×5+7×5 = 50) dimension vector. So a 230 dimension vector is used as an input to their classifier which is a SVM classifier. They obtained accuracy and MCC after filtering the final prediction of 79.2% and 0.48 respectively. The same features that are used as input to predict beta-turn/non-turn structure are supplied to SVMs classifiers to recognize the different types of beta-turns and the accuracies obtained are 78.6, 87.4, 71.5, 71.1, 97.6 for types I, II, IV, VIII, NS respectively, where the other types are combined in type NS.

PSSMs, Predicted Secondary Structures and Predicted Shape String

Tang *et al*. (2011) and our own method (Elbashir *et al*., 2013a) utilized shape strings together with PSSM and predicted secondary structure to predict beta-turns in protein both Tang *et al*. and our method are SVM methods. The shape strings can be predicted from a predictor constructed based on structural alignment approach. The eight states S, R, U, V, K, A, T and G represents the shape strings of a protein. A detailed information about protein structure including random coil in which beta-turn is located can be provided by shape strings. This can make them as important component that can be used to predict beta- turns. Both of the methods used protein shape string and its Profile Prediction Server (DSP) (Sun *et al*. 2012) to obtain the predicated shape strings. The eight states of the shape string are encoded using the binary encoding schema. In parts of proteins sequence there can be a location where the j and ψ angles are undefined, or the structure determination for it may be unknown. For these specific parts the DSP server defines additional shape N. an example for the binary encoding schema where the shape is S is (1 0 0 0 0 0 0 0 0) and where the shape is N is (0 0 0 0 0 0 0 0 1). In our method (Elbashir *et al*., 2013a) we

used a cluster model to deal with the imbalance problem in predicting beta-turns. In the cluster model the non-b-turns set is divided into a three subsets by k-means clustering algorithm and then three SVMs are used, each of them used one cluster of the non-beta-turns against the beta-turns and then a logistic regression model, modeled using fractional polynomial is used to aggregate the results of the three SVMs. The accuracy and MCC achieved using our own method are 87.37 and 0.67 respectively.

PSSM, Predicted Shape String and Predicted Protein Block

In addition to PSSM and predicted shape string, Lan Anh T. Nguyen *et al*., added predicted protein block (de Brevern *et al*., 2000; de Brevern, 2005; Joseph *et al*., 2010), which they obtained from the web site of PB-kPRED. Sixteen pentapeptide motifs with labels A, B, C, D, E, F, G, H, I, J, K, L, M, N, O and P determine the structural alphabet of the predicted protein block (de Brevern *et al*., 2000; de Brevern 2005; Tyagi *et al*., 2006). To deal with imbalance problem in predicting beta-turns (25% turn vs 75% non-turn), Lan Anh T. Nguyen *et al* used oversampling technique with SVM as a base classifier. They used a window size of 9 on the PSSM, predicted shape string and predicted protein blocks and obtained accuracy and MCC of 87.48 and 0.66 respectively. For bet-turns types prediction they combined types VIa1, VIa2 and VIb in one type named VI that is because types VIa1, VIa2 and VIb are rare (Chou, 2000). Lan Anh T. Nguyen *et al*. obtained an accuracy of 93.45, 99.28, 97.90, 99.44, 90.18, 98.07, 90.18 on types I, I', II, II', IV, VI, VIII respectively. And MCC of 0.61, 0.75, 0.75, 0.64, 0.38, 0.14, 0.30.

Discussion

The methods that are applied on beta-turns prediction and their types use different proteins sequence features. These features include the basic sequence information, PSSMs, predicted secondary structure, predicted dihedral angle and predicted surface accessibility and the predicted shape string. Table 2 summarizes the results of predicting beta-turns that are obtained by the different prediction methods together with the features they used, while Table 3 summarizes the results of predicting beta-turns types that are obtained by the different prediction methods and the features they used. PSSMs are proved to be having a significant contribution in accuracy of beta-turns prediction compared to the basic sequence information. Therefore, PSSMs are used in almost all of the most successful methods that are constructed for beta-turns predictions. In most of the successful beta-turns predictions methods, PSSMs are generated using several rounds of the PSI-BLAST program (Altschul *et al*., 1997) against National Center for Biotechnology Information (NCBI) nonredundant (nr) database. A window based approach is used to compose the input vector from the PSSM. Some of the methods used a window size of 9 whereas most of the methods used a window size of 7. Figure 1 depicts the use of window size of 7 on a PSSM.

Most of the successful methods used an equation to scale the value of the PSSMs to a range between 0 and 1. Predicted secondary structures are combined with PSSMs to enhance the prediction accuracy. This combination is organized differently in the researches.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
G	1	-3	-2	-3	-3	-3	-3	6	-3	-5	-5	-3	-4	-4	-3	-1	-2	-4	-4	-4
G	-1	-2	0	2	-4	1	4	4	-2	-4	-4	-1	-3	-4	0	1	0	-4	-4	-4
G	-1	-2	-1	-1	-4	0	2	4	-3	-4	-3	0	-3	-4	3	0	0	-4	-4	-3
I	0	0	-1	0	-2	-1	1	0	-2	2	0	-1	0	0	2	-1	-1	-3	-2	1
L	-1	0	-1	-1	-1	1	1	-1	-1	0	1	0	0	1	-1	-1	-1	-1	2	0
D	-1	-1	0	3	0	0	4	0	-2	-4	-4	3	-3	-4	-3	0	0	-4	-3	-3
S	0	-1	1	0	-3	-1	2	2	-2	-3	-3	2	-3	-4	-1	2	1	-4	-3	-2
M	1	-1	0	0	-3	1	3	0	-2	-2	-1	0	-1	-3	0	2	0	-4	-3	0
V	1	-1	0	-1	0	0	0	0	-1	0	0	0	0	-2	0	1	0	-3	-2	1
E	-2	-2	-1	2	-5	3	5	-3	-2	-3	-4	1	-4	-5	0	-2	-3	-5	-3	-1
K	-1	1	1	-1	-4	2	2	-3	2	-2	-3	3	-3	-3	-2	1	1	-4	1	0
L	-3	4	-5	-5	4	-2	-4	-5	-4	-1	4	-3	0	-3	-4	-4	-3	3	-3	0
G	-4	-6	-4	-5	-6	-6	-6	8	-6	-8	-8	-5	-7	-7	-6	-4	-5	-6	-7	-7
K	-2	5	-1	1	-1	2	2	-4	-2	-3	-4	3	0	-5	-4	-1	0	-5	-4	-3
L	-4	-5	-6	-6	-4	-5	-6	-7	-6	5	5	-5	1	-3	-6	-5	-4	-5	-4	1
Q	-3	-1	2	-2	3	4	1	-4	5	-3	0	-1	0	-1	-4	-1	-2	3	0	-2

Fig. 1: A window size of 7 on a PSSM

DOI: 10.3844/jcssp.2019.1331.1340

Table 2: The results of predicting beta-turns that are obtained by the different prediction methods and the features used

Prediction method	Features used	Accuracy	MCC
Shepherd *et al*. (1999)	Basic Sequence information and reliability indices	74.9%	0.35
Kirschner and Frishman (2008)	PSSMs	77.9%	0.45
Zhang *et al*. (2005)	PSSMs and predicted secondary structure	74.8 %	0.41
Kaur and Raghava (2003)	PSSMs and predicted secondary structure	75.5%	0.43
Zheng and Kurgan (2008)	PSSMs and predicted secondary structure	80.7%	0.50
Elbashir *et al*. (2013a)	PSSMs and predicted secondary structure	82.87%	0.56
Petersen *et al*. (2010)	PSSMs, Predicted secondary structure and surface ccessibility	78.2%	0.50
Kountouris and Hirst (2010)	PSSMs, Predicted backbone dihedral angle and secondary structure.	79.2%	0.48
Tang *et al*. (2011)	PSSMs, Predicted secondary structures and predicted shape string	87.2%	0.66
Elbashir *et al*. (2013a)	PSSMs, Predicted secondary structures and predicted shape string	87.37	0.67
Nguyen *et al*. (2014)	PSSM, predicted shape string and predicted protein block.	87.48	0.66

Table 3: The results of predicting beta-turns types that are obtained by the different prediction methods and the features used

Prediction method	Features used	Beta-turns Type	Accuracy	MCC
Shepherd *et al*. (1999)	Basic Sequence information and reliability indices	I	91.2%	0.219
		II	95.5%	0.253
		VIII	95.7%	0.062
		IV	96.8%	0.033
Kirschner and Frishman (2008)	PSSM	I (Maximizing MCC)	82.5%	0.317
		I (tradeoff between the MCC and the accuracy)	85.4%	0.314
		II (Maximizing MCC)	96.2%	0.339
		II (tradeoff between the MCC and the accuracy)	96.2%	0.339
		VIII (Maximizing MCC)	53.4%	0.109
		VIII (tradeoff between the MCC and the accuracy)	93.0%	0.076
		IV (Maximizing MCC)	72.3%	0.236
		IV (tradeoff between the MCC and the accuracy)	85.2%	0.196
		I' (Maximizing MCC)	98.8%	0.356
		I' (tradeoff between the MCC and the accuracy)	98.8%	0.356
		II' (Maximizing MCC)	98.6%	0.137
		II' (tradeoff between the MCC and the accuracy)	98.6%	0.137
Petersen *et al*. (2010)	PSSMs, Predicted secondary structure and surface accessibility	I,	N/A	0.36
		I'	N/A	0.23
		II	N/A	0.31
		II'	N/A	0.16
		IV	N/A	0.27
		VIII	N/A	0.16
Kountouris and Hirst (2010)	PSSMs, Predicted backbone dihedral angle and secondary structure	I	78.6%	N/A
		II	87.4%	N/A
		IV	71.5%	N/A
		VIII	71.1%	N/A
Nguyen *et al*. (2014)	PSSM, predicted shape string and predicted protein block.	I	93.45%	0.61
		I'	99.28%	0.75
		II	97.90%	0.75
		II'	99.44%	0.64
		IV	90.18%	0.38
		VI	98.07%	0.14
		VIII	90.18%	0.30

Some of the researcher used a sliding window on the PSSMs only and then the three state secondary structures are attached to the feature vector, where others used a sliding window on both PSSMs and predicted secondary structures. In our own work (Elbashir *et al*.. 2013), we tried both of these organizations and we found that using sliding window on Both PSSMs and predicted secondary structures gives better classification results. The method that is

DOI: 10.3844/jcssp.2019.1331.1340

constructed by Zheng and Kurgan (2008) was the first method to predict beta-turns at over 80% accuracy. This method used four protein secondary structure prediction methods to extract several secondary structure information and then combine these information with the PSSMs in different organization. Figure 2 to 5 show this combinations. Figure 2 depicts the secondary structure features that are extracted from the four secondary structure prediction methods (PSIPRED, JNET, TRANSEC and PROTEUS). Figure 3 shows the confidence value of the central residue for each of the prediction method, which are used as features in addition with the secondary structure features. Figure 4 shows the binary values representing a specific arrangement of the secondary structure predicted with the four prediction methods for the central and the two adjacent residues, in the figure TRANSEC is shown as an example of the prediction methods. Figure 5 shows the features that are taken from the ratio between the number of residues in a given secondary structures and the window size for each of the prediction method.

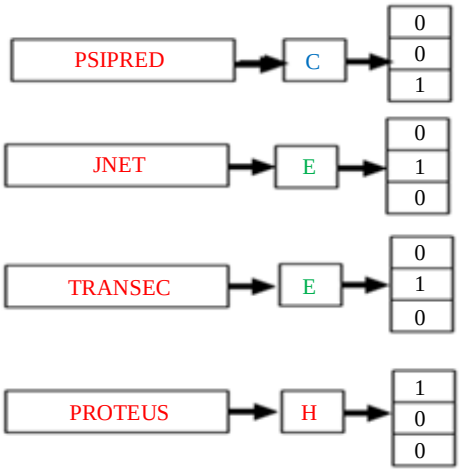

Fig. 2: Secondary structure prediction for each of the prediction method

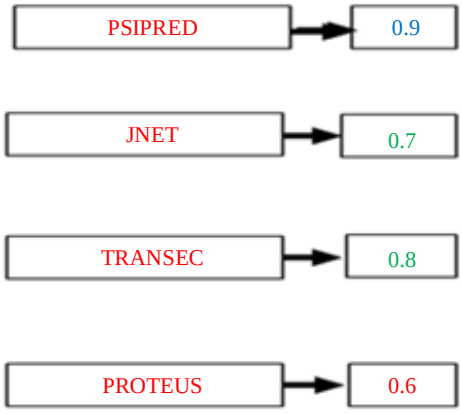

Fig. 3: The confidence value of the central residue for each of the prediction method

A great leap in the prediction of beta-turns was obtained after adding the predicted shape string of the protein to the PSSM and the predicted secondary strictures to form the input features. The accuracy and The MCC that were obtained after adding the predicted shape string of the protein are more the 87% and 55% respectively. Figure 6 shows how the PSSMs, predicted secondary structures and the predicted shape string features are represented.

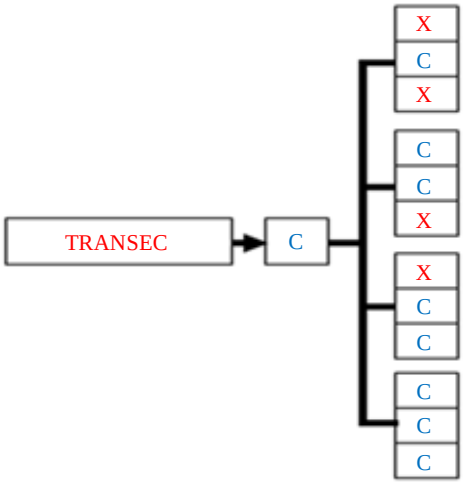

Fig. 4: Binary values representing a specific arrangement of the secondary structure predicted with the four prediction methods for the central and the two adjacent residues (the figures shows the prediction using TRANSEC)

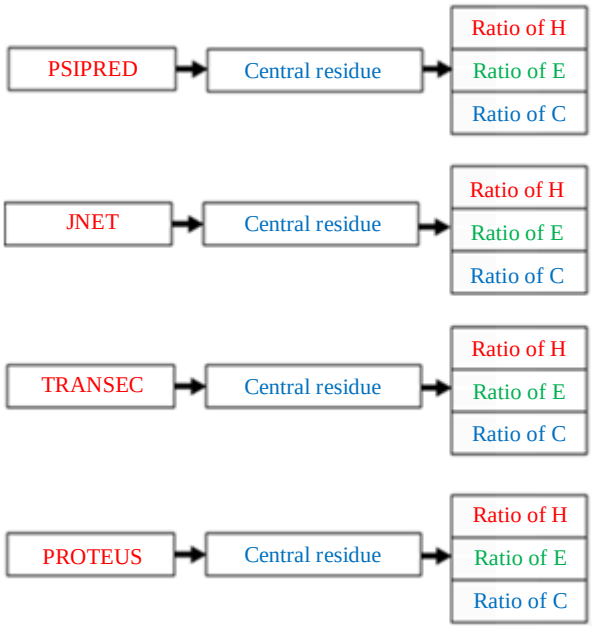

Fig. 5: The ratio between the number of residues in a given secondary structures and the window size for each of the prediction method

DOI: 10.3844/jcssp.2019.1331.1340

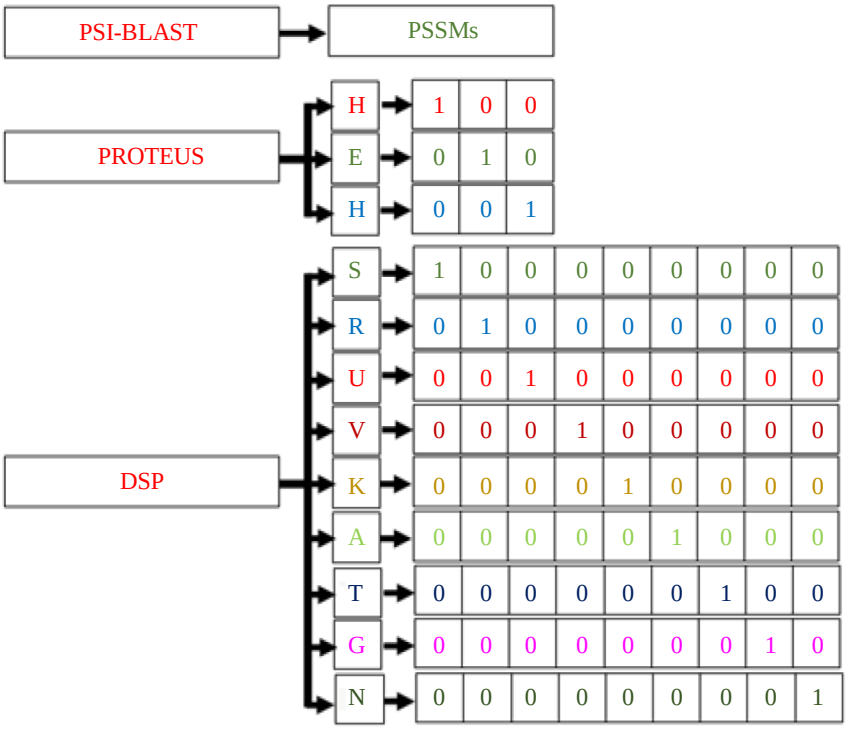

Fig. 6: PSSMs created using PSI-BLAST, PSS predicted using Proteus and the shape strings predicted using the protein shape string and its profile prediction server (DSP)

de Brevern (2016) extended the classical classification (Venkatachalam, 1968; Richardson, 1981; Chan *et al*., 1993; Hutchinson and Thornton, 1996) of beta-turns types by adding additional beta-turn types. Shapovalov *et al*. (2019) defined new 18 turn types. These new added beta-turns types should be considered in future researches that predict beta-turns types in proteins. Deep learning, which is a rapidly growing research area and many NN architectures are designed to implement it awaits wide applications in bioinformatics (Zhang and Rajapakse, 2009). Its NN architectures consist of multiple nonlinear layers and there are several types of these architectures according to the input characteristic and the objectives for which it is designed (Liu *et al*., 2017). Deep learning can make a breakthrough in beta-turns prediction, because the features will be automatically created by the NN when it learns, but this does not mean that obtaining features and pre-process it is totally irrelevant. Extracted features such as PSSMs and predicted secondary structures can be used as an input for deep learning algorithms to ease difficulties from complex biological data and improve performance (Zhang and Rajapakse 2009).

Conclusion

The protein secondary structure is considered to be the base of analyzing the functional properties of the protein. These functional properties depend on the protein three-dimensional structure. The beta-turns is the most important part of protein secondary structure, therefore their prediction is crucial for the advancement in protein folding and drug design. The methods that are designed for beta-turns prediction used different kinds of features. The most used features in the recent prediction methods is the PSSMs. Although beta-turns itself is one of the secondary structures types, the predicted secondary structures obtained using different prediction servers are added to the PSSMs to form the input vector for prediction methods. These prediction methods used different PSSMs and predicted secondary structures organization and some of them used another secondary structure information combined with PSSMs and predicted secondary structures to form the input vector. Some methods used other features such as surface accessibility and predicted backbone dihedral angle combined with predicted secondary structures and PSSMs. The state of the art methods that obtained the highest classification performance have used either predicted shape string in a combination with PSSMs and predicted secondary structure or predicted protein block in a combination with predicted shape strings and PSSMs.

Acknowledgement

We would like to acknowledge Jouf University for all the support that it provides. This work is supported by Jouf University under Research Project No. 39/751.

Author's Contributions

MKE reviewed the literature and drafted the manuscript.

Ethics

This article is original and contains unpublished materials. No ethical issues involved.

References

Adamczak, R., A. Porollo and J. Meller, 2005. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins, 59: 467-475. DOI: 10.1002/prot.20441

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang and Z. Zhang *et al*., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25: 3389-3402. DOI: 10.1093/nar/25.17.3389

Bryson, K., L.J. McGuffin, R.L. Marsden, J.J. Ward and J.S. Sodhi *et al*., 2005. Protein structure prediction servers at University College London. Nucleic Acids Res., 33: W36-38. DOI: 10.1093/nar/gki410

Chan, A.W., E.G. Hutchinson, D. Harris and J.M. Thornton, 1993. Identification, classification and analysis of beta-bulges in proteins. Protein Sci., 2: 1574-1590. DOI: 10.1002/pro.5560021004

Chou, K.C., 1997. Prediction of beta-turns. J. Pept. Res., 49: 120-144. DOI: 10.1111/j.1399-3011.1997.tb00608.x

Chou, K.C., 2000. Prediction of tight turns and their types in proteins. Anal. Biochem., 286: 1-16. DOI: 10.1006/abio.2000.4757

Chou, K.C. and J.R. Blinn, 1997. Classification and prediction of beta-turn types. J. Protein Chem., 16: 575-595. DOI: 10.1023/A:1026366706677

Chou, P.Y. and G.D. Fasman, 1974. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry, 13: 211-222. DOI: 10.1021/bi00699a001

Cuff, J.A. and G.J. Barton, 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40: 502-511. DOI: 10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q

de Brevern, A.G., 2005. New assessment of a structural alphabet. In Silico Biol., 5: 283-289.

de Brevern, A.G., 2016. Extension of the classical classification of beta-turns. Sci. Rep., 6: 33191-33191. DOI: 10.1038/srep33191

de Brevern, A.G., C. Etchebest and S. Hazout, 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins, 41: 271-287. DOI: 10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z

Elbashir, M., J. Wang, F.X. Wu and L. Wang, 2013a. Predicting beta-turns in proteins using support vector machines with fractional polynomials. Proteome Sci., 11: S5-S5. DOI: 10.1186/1477-5956-11-S1-S5

Elbashir, M.K., Y. Sheng, J. Wang, F. Wu and M. Li, 2013b. Predicting beta-turns in protein using kernel logistic regression. Biomed. Res. Int., 2013: 870372-870372. DOI: 10.1155/2013/870372

Fuchs, P.F. and A.J. Alix, 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins, 59: 828-839. DOI: 10.1002/prot.20461

Guruprasad, K. and S. Rajkumar, 2000. Beta-and gamma-turns in proteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials. J. Biosci., 25: 143-156. DOI: 10.1007/BF03404909

Hu, X. and Q. Li, 2008. Using support vector machine to predict beta- and gamma-turns in proteins. J. Comput. Chem., 29: 1867-1875. DOI: 10.1002/jcc.20929

Hutchinson, E.G. and J.M. Thornton, 1994. A revised set of potentials for beta-turn formation in proteins. Protein Sci., 3: 2207-2216. DOI: 10.1002/pro.5560031206

Hutchinson, E.G. and J.M. Thornton, 1996. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci., 5: 212-220. DOI: 10.1002/pro.5560050204

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292: 195-202. DOI: 10.1006/jmbi.1999.3091

Joseph, A.P., G. Agarwal, S. Mahajan, J.C. Gelly and L.S. Swapna *et al*., 2010. A short survey on protein blocks. Biophys. Rev., 2: 137-147. DOI: 10.1007/s12551-010-0036-1

Kaur, H. and G.P. Raghava, 2003. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci., 12: 627-634. DOI: 10.1110/ps.0228903

Kee, K.S. and S.D. Jois, 2003. Design of beta-turn based therapeutic agents. Curr. Pharm. Des., 9: 1209-1224. DOI: 10.2174/1381612033454900

Kim, S., 2004. Protein beta-turn prediction using nearest-neighbor method. Bioinformatics, 20: 40-44.

Kirschner, A. and D. Frishman, 2008. Prediction of β-turns and β-turn types by a novel bidirectional Elman-type recurrent neural network with multiple output layers (MOLEBRNN). Gene, 422: 22-29. DOI: 10.1016/j.gene.2008.06.008

Kountouris, P. and J.D. Hirst, 2009. Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformat., 10: 437-437. DOI: 10.1186/1471-2105-10-437

Kountouris, P. and J.D. Hirst, 2010. Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures. BMC Bioinformat., 11: 407-407. DOI: 10.1186/1471-2105-11-407

Elman, J., 1990. Finding structure in time. Cognitive Sci., 14: 179-211. DOI: 10.1207/s15516709cog1402_1

Liu, L., Y. Fang, M. Li and C. Wang, 2009. Prediction of beta-turn in protein using E-SSpred and support vector machine. Protein J., 28: 175-181. DOI: 10.1007/s10930-009-9181-4

Liu, W., Z. Wang, X. Liu, N. Zeng and Y. Liu *et al*., 2017. A survey of deep neural network architectures and their applications. Neurocomputing, 234: 11-26. DOI: 10.1016/j.neucom.2016.12.038

McGregor, M.J., T.P. Flores and M.J. Sternberg, 1989. Prediction of beta-turns in proteins using neural networks. Protein Eng., 2: 521-526. DOI: 10.1093/protein/2.7.521

McGuffin, L.J., K. Bryson and D.T. Jones, 2000. The PSIPRED protein structure prediction server. Bioinformatics, 16: 404-405. DOI: 10.1093/bioinformatics/16.4.404

Montavon, G., G. Orr and K.R. Müller, 2012. Neural Networks: Tricks of the Trade. 2nd Edn., Springer, New York, ISBN-10: 3642352898, pp: 769.

Montgomerie, S., S. Sundararaj, W.J. Gallin and D.S. Wishart, 2006. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformat., 7: 301-301. DOI: 10.1186/1471-2105-7-301

Nguyen, L.A.T., X.T. Dang, T.K.T. Le, T. Saethang and V.A. Tran *et al*., 2014. Predicting βeta-turns and βeta-turn types using a novel over-sampling approach. J. Biomed. Sci. Eng., 7: 927-940. DOI: 10.4236/jbise.2014.711090

Petersen, B., C. Lundegaard and T.N. Petersen, 2010. NetTurnP--neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. PLoS One, 5: e15079-e15079. DOI: 10.1371/journal.pone.0015079

Petersen, B., T.N. Petersen, P. Andersen, M. Nielsen and C. Lundegaard, 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol., 9: 51-51. DOI: 10.1186/1472-6807-9-51

Pham, T.H., K. Satou and T.B. Ho, 2003. Prediction and analysis of beta-turns in proteins by support vector machine. Genome Inform., 14: 196-205.

Qian, N. and T.J. Sejnowski, 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202: 865-884. DOI: 10.1016/0022-2836(88)90564-5

Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv. Protein Chem., 34: 167-339-339. DOI: 10.1016/S0065-3233(08)60520-3

Rost, B. and C. Sander, 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232: 584-599. DOI: 10.1006/jmbi.1993.1413

Shapovalov, M., S. Vucetic and R.L. Dunbrack, Jr., 2019. A new clustering and nomenclature for beta turns derived from high-resolution protein structures. PLOS Computat. Biol., 15: e1006844-e1006844. DOI: 10.1371/journal.pcbi.1006844

Shepherd, A.J., D. Gorse and J.M. Thornton, 1999. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci., 8: 1045-1055. DOI: 10.1110/ps.8.5.1045

Sun, J., S. Tang, W. Xiong, P. Cong and T. Li, 2012. DSP: A protein shape string and its profile prediction server. Nucleic Acids Res., 40: W298-302. DOI: 10.1093/nar/gks361

Tang, Z., T. Li, R. Liu, W. Xiong and J. Sun *et al*., 2011. Improving the performance of beta-turn prediction using predicted shape strings and a two-layer support vector machine model. BMC Bioinformat., 12: 283-283. DOI: 10.1186/1471-2105-12-283

Tyagi, M., P. Sharma, C.S. Swamy, F. Cadet and N. Srinivasan *et al*., 2006. Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res., 34: W119-123. DOI: 10.1093/nar/gkl199

Venkatachalam, C.M., 1968. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers, 6: 1425-1436. DOI: 10.1002/bip.1968.360061006

Wilmot, C.M. and J.M. Thornton, 1988. Analysis and prediction of the different types of beta-turn in proteins. J. Mol. Biol., 203: 221-232. DOI: 10.1016/0022-2836(88)90103-9

Wilmot, C.M. and J.M. Thornton, 1990. Beta-turns and their distortions: A proposed new nomenclature. Protein Eng., 3: 479-493. DOI: 10.1093/protein/3.6.479

Zhang, C.T. and K.C. Chou, 1997. Prediction of β- turns in proteins by 1-4 and 2-3 correlation model. Biopolymers, 41: 673-702. DOI: 10.1002/(SICI)1097-0282(199705)41:6<673::AID-BIP7>3.0.CO;2-N

Zhang, Q., S. Yoon and W.J. Welsh, 2005. Improved method for predicting beta-turn using support vector machine. Bioinformatics, 21: 2370-2374. DOI: 10.1093/bioinformatics/bti358

Zhang, Y. and J.C. Rajapakse, 2009. Machine Learning in Bioinformatics. 1st Edn., Wiley, Hoboken, ISBN-10: 0470397411, pp: 400.

Zheng, C. and L. Kurgan, 2008. Prediction of beta- turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformat., 9: 430-430. DOI: 10.1186/1471-2105-9-430