PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(5): 472~482
www.pibb.ac.cn

Using Homology Information From PDB to Improve The Accuracy of Protein β-turn Prediction by NetTurnP*

QIAN Gang[1,2], WANG Hai-Yan[3], YUAN Zhe-Ming[1,2]**

([1] *Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha* 410128, China;
[2] *Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Changsha* 410128, China;
[3] *Department of Statistics, Kansas State University, Manhattan, Kansas* 66506, USA)

Abstract β-Turn is a secondary protein structure type that is important in protein folding, protein stability and molecular recognition processes. To date, various methods have been put forward to predict β-turns, but none of them have tried directly to map the structures of pre-existing homologues from structural databases like RCSB PDB to the protein to be predicted. Given the large size of PDB (>70 000 structures), it is actually of high possibility to find a structural homologue for a newly identified sequence. In this work, we present a new method that predicts β-turns by combining homology information extracted from PDB with the results predicted by NetTurnP. Two datasets, the golden set BT426 and the self-constructed dataset EVA937, are used to assess our method. For each sequence in both datasets, only homologues deposited earlier than the sequence in PDB are employed. We have achieved Matthews correlation coefficients (MCCs) of 0.56, 0.52 respectively, which are higher than those obtained by NetTurnP alone of 0.50, 0.46, and the prediction accuracies (Q_{total}) obtained using our method are 81.4% and 80.4% separately, while NetTurnP alone achieves 78.2% and 77.3%. The results confirm that combining the homology information with state-of-the-art β-turn predictors like NetTurnP can significantly improve the prediction accuracy. A Java program called BTMapping has been written to implement our method, which is freely available at http://www.bio530.weebly.com together with the related datasets.

Key words β-turn prediction, homology information, PDB, NetTurnP, BTMapping
DOI: 10.3724/SP.J.1206.2011.00370

Information about secondary protein structures is useful for a wide range of applications including prediction of solvent accessibility[1], fold type[2], folding rate[3], β-turns[4], α-turns[5], contact order[6], tertiary structure[7], and fold recognition[8]. Therefore, its prediction has been an area of intense research over the past three decades. The secondary structure of a protein can be classified as local structural elements of α-helices, β-strands and coil regions. β-Turns are actually ordered local structures of coil regions. On average, about 25% of residues in protein structures form β-turns[9], so they are one of the most abundant secondary structures. A β-turn consists of four consecutive residue which are not in an α-helix, and the distance between the Cα-atoms i, i+3 is less than 7 Å. β-Turns can be further classified into nine subtypes[10], according to the dihedral angles between amino acid residues i+1 and i+2.

β-Turns play many significant roles in the structure and function of protein and peptide. Because of their four-residue reversals in protein, β-turn

*This work was supported by grants from The Science Foundation for Distinguished Young Scholars of Hunan Province, China (10JJ1005), The Research Fund for The Doctoral Program of Higher Education of China (200805370002).

**Corresponding author.
Tel: 86-731-84613956, E-mail: zhmyuan@sina.com

Received: August 11, 2011 Accepted: November 4, 2011

formation is an important step in the process of protein folding [11], while improved β-turn sequences can improve protein stability[12–13]. Additionally, β-turns are crucial components of β-hairpins and anti-parallel β-sheets, whose prediction has recently attracted interest[14–15]. Moreover, β-turns tend to be more solvent exposed than buried and thus they are often related to molecular recognition and modeling interactions between peptide substrates and receptors [16]. In recent years, research interest has been aroused in mimicking β-turns for the synthesis of medicines [17–18] and nucleating β-sheet folding[19].

Due to the importance of β-turns in biology, many β-turn prediction methods have been proposed so far. They can be divided into statistical methods and machine learning techniques. Table 1 lists main β-turn prediction methods and their performance of 7-fold cross-validation on the golden dataset BT426. Statistical methods utilize probabilities computed using information concerning preferences of individual amino acid types at each position in β -turns. It is shown in Table 1 that the statistical methods have poorer performance, and among them only a recent method called COUDES [4] obtains a MCC of 0.42, while others [20–24] only obtain MCCs in the range of 0.17 ～0.33. Compared with statistical methods, machine learning methods have better performance. The most accurate β-turn predictors today utilize machine learning techniques. Neural network was the first used to predict β-turns as a machine learning method though the first version only reached MCC accurary of 0.2 [25]. Since then, neural networks have been frequently used and improved for β-turn prediction[26–30]. NetTurnP[30] is the latest neural network β-turn predictor that uses two layers of neural networks and achieves a MCC of 0.49, which is the highest reported performance as a *de novo* (sequence-based) predictor on a two-class prediction of β-turn and non-β-turn. Other machine learning techniques introduced to β-turn prediction include k-nearst-neighbor [31] that reached a MCC of 0.40 and support vector machines [8, 32–37] (SVM) achieving MCCs in the range of 0.44～0.48. Different from earlier versions of SVM predictors is a recent two-layer SVM predictor ShapeString_Pred[38], which utilizes predicted secondary structures and predicted shape strings as input features. September 2010 release of PDB homologue information was indirectly used in ShapeString_Pred since the predicted shape strings were actually derived from PDB homologues using a web server and part of the secondary structures were predicted by Proteus which also used PDB homologues. The dataset BT426, however, was created in 2000[39]. As PDB database from a later date contains more homologues to a query sequence that can be found through BLAST, it is of higher possibility to find optimal homologues for structure mapping to reach a better prediction accuracy. ShapeString_Pred achieved a MCC of 0.66 for the BT426. With the same release of the PDB information, the method in this article had a MCC of 0.7152. Further discuss of how the prediction accuracy changes with the date of PDB release is deferred to Section 2.4.

For a newly identified protein sequence, however, only structures deposited earlier in the PDB can actually be employed for structure mapping. Though quite a few methods are available to predict β-turns, none of them are able to directly assign the structures of deposited homologues in structural databases to the target protein sequence. However, similar strategy has been used for the prediction of protein secondary structure and improved the classification accuracy of the basic local structural elements of α-helices, β-strands and coil regions[40]. It has been reported that less than 3% of new protein structures deposited into the PDB have a totally novel fold[41], and nearly 3/4 of newly deposited PDB structures have sequence identities greater than 25% to a pre-existing structure[40]. In this work, we developed a method to improve β-turn prediction by combining the direct homology information extracted from PDB with NetTurnP[30]. For each query sequence to be predicted, its homologues were obtained by using BLAST [42] against a recent release of PDB sequences. Then a multiple sequence alignment was conducted by ClustalW[43–44] to align the query sequence with all its homologues. Finally, a mapping technique was used to map the structure (β-turn or not β-turn) of the optimal homologue to the query sequence for each position in the alignment. The prediction accuracy of the new method under different sequence identity levels from BLAST hits of the query sequences was investigated. A Java program called BTMapping was written to complete all the above mentioned processes, which is accessible at http://www.bio530.weebly.com.

Table 1 Summary of different β-turn prediction methods on BT426 dataset using a 7-fold cross-validation

Type	β-Turn predictor	Measure			
		MCC	Q_{total}/%	Q_{pre}/%	Q_{obs}/%
SVM	ShapeString_Pred[38]	0.66	87.2	73.8	75.9
	DEBT[37]	0.48	79.2	54.8	70.1
	Zheng and Kurgan[34]	0.47	80.9	62.7	55.6
	Hu and Li[35]	0.47	79.8	55.6	68.9
	Zhang *et al.*[33]	0.45	77.3	53.1	67.0
	BTSVM[32]	0.45	78.7	56.0	62.0
	Liu *et al.*[36]	0.44	80.9	63.6	49.2
NN	MOLEBRNN[28]	0.45	77.9	53.9	66.0
	BETAPRED2[27]	0.43	75.5	49.8	72.3
	BTPRED[26]	0.35	74.4	48.3	57.3
	NetTurnP[30]	0.50	78.2	54.4	75.6
	NetTurnP-tweak[30]	0.48	82.1	68.8	50.9
	NetTurnP 7-fold[30]	0.49	78.1	54.4	74.2
KNN	Kim[31]	0.40	75.0	46.5	66.7
ST	1-4&2-3 correlation model[21, 45]	0.17	59.1	32.4	61.9
	GORBTURN[20, 45]	0.19	70.5	39.3	37.3
	Chou-Fasman[22, 45]	0.26	65.2	37.6	63.5
	Thornton[23, 45]	0.23	68.0	38.6	52.4
	Sequence coupled model[24, 45]	0.33	72.2	45.0	60.0
	COUDES[4]	0.42	74.8	48.8	69.9

SVM: support vector machine; NN: neural network; KNN: k-nearest neighbor; ST: statistical. In the table, there are three rows for NetTurnP. From the top down, NetTurnP is referring to the performace of direct prediction on BT426 that was treated as an independent test set, NetTurnP-tweak is the approach that was tweaked for best Q_{total} performance, and NetTurnP 7-fold is referring to a 7-fold cross-validation performed on the BT426 dataset.

1 Materials and methods

1.1 Datasets

Two datasets, BT426 and a newly-constructed EVA937, were used to evaluate the performance of our method, while the third new dataset EVA300 was used for parameter optimization. BT426 was developed by Guruprasad and Rajkumar in 2000[39]. In BT426, there are 426 protein sequences, with the average length 223.6 residues, the β-turns were assigned using PROMOTIF[46]. The pairwise sequence identity between any two protein chains is below 25%. The structure was determined by X-ray crystallography with at least 2.0 Å resolution, and each chain contains at least one β-turn. We downloaded the BT426 dataset following the link given by one recent paper that introduces predictior BTNpred [34]. In this version of BT426 dataset,one sequence named "1adoa" was corrected for its inconsistency with the corresponding sequence in PDB. EVA937 was newly constructed. Since β-turn prediction is closely related to protein secondary structure prediction, a dataset used to evaluate a secondary structure predictor Proteus was revised to make it suitable for β-turn prediction, and this dataset was actually derived from EVA[47] that continously and automatically analyses protein structure prediction servers in "real time". The dataset downloaded from the Proteus website initially have 1 774 PDB protein IDs. According to the ID, the corresponding pdb file can be downloaded, and then the β-turns can be assigned by PROMOTIF [46]. For all 1 774 IDs, the sequence represented by the ID that meets the following several rules was then selected: (1) The amino acid sequence contains at least 15 residues; (2) There are no residues of unknown amino acid type in the sequence; (3) The atom coordinates from the pdb file of the ID are continuous and complete for every amino acid. Finally, 1 237 proteins were selected. No pair of these proteins has more than 33% identical residues over more than 100 residues aligned. 300 proteins were randomly selected as EVA300 to optimize the parameters of our method, and the remaining 937 proteins are the EVA937 used to test the performance of our method. The average length of sequences in EVA937 is 194.4 residues.

1.2 Using BLAST against PDB to find homologues

The latest 2.2.25 linux version of BLAST program was downloaded from NCBI's ftp site at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST. A FASTA file containing all PDB sequences by April 5, 2011, was downloaded from PDB's structure download page at http://www.pdb.org/pdb/download/download.do. All proteins were selected and filtered by CD-HIT[48] utility at 95% sequence identity threshold to accelerate the BLAST process, which means all sequences with > 95% sequence identity to any other sequence were removed. When we were using any one of the three datasets (BT426, EVA300, EVA937), all sequences contained in the dataset were removed from the filtered FASTA file. The formatdb tool in the BLAST program was further used to format the file to the database files recognized by BLAST. For each query sequence of the three datasets, its homologues were found by BLAST against the filtered set of PDB sequences. The BLAST command we used is "blastall -p blastp -d pdb -i queryFile -o outputFile -e 1e-3 -F F -a 4", and the expectation value following the tag "-e" is 10^{-3}, which is enough to include all the hits we need. Only hits with more than 25% sequence identity to the query sequence were treated as homologues used for structure mapping, and only hits deposited earlier in the PDB than the query sequence were selected for processing the three datasets used in this paper, which is similar to the situation when we try to predict the β-turns of a newly identified sequence. A parameter called "byDate" in our program is responsible for controlling the selection of the hits deposited earlier than a specified date. For the hits with more than 100 residues, this process is further controlled by a parameter called $I_{\max}$, which is the maximum sequence identity of the hits to be selected to the query sequence.

1.3 Using ClustalW to align the query and all corresponding hits

ClustalW [43-44] is a program widely used for multiple sequence alignment, which was used to align the query sequence from the evaluation dataset with all its homology hits from PDB. The current 2.1 version of ClustalW can be downloaded from http://www.clustal.org/.

1.4 Mapping the structure of homologues to the query sequence

For each query sequence, the prediction result of NetTurnP can be obtained from a website at http://www.cbs.dtu.dk/services/NetTurnP/. According to the multiple sequence alignment by ClustalW, a strategy of sliding a 7 residue window over the alignment was used to map the structures of homologues to the query sequence. The same window size was adopted by Proteus [40] using similar mapping strategy to predict protein secondary structures. Each non-gap (i. e. not "-") residue of the query sequence marks a column in the whole alignment, centered around which a 7 residue window was opened. In this window, the central non-gap residue of each aligned homologue was assigned a probability value P_{h}, which is defined as follows:

$$P_{\mathrm{h}}=0.5+0.5 \cdot I'_{\mathrm{g}} \cdot I_{\mathrm{l}} \tag{1}$$

$$I'_{\mathrm{g}}=\begin{cases} 1 & \text{if } I_{\mathrm{g}}>T_{\mathrm{g}} \\ I_{\mathrm{g}} & \text{if } I_{\mathrm{g}}\leqslant T_{\mathrm{g}} \end{cases} \tag{2}$$

$$I_{\mathrm{l}}=N_{\mathrm{s}}/L_{\mathrm{w}} \tag{3}$$

These equations are defined based on the following analysis. First, we assume the probability (i. e. P_{h}) of assigning the type of the central residue of a homologue over the window to the aligned residue of the query sequence is 0.5. P_{h} can be increased depending on the sequence identity of the homologue to the query sequence over the whole sequence and over the window. The higher both identities are, the higher P_{h} would be. When the sequence identity over the whole sequence exceeds a threshold, P_{h} mainly depends on the sequence identity over the window. Among these equations, I_{g} stands for the sequence identity of the homologue to the query sequence over the whole sequence using BLAST. I'_{g} is derived from I_{g} using a threshold T_{g}. I_{l} represents the sequence identity of the homologue to the query sequence over the window, which is calculated by dividing the number of identical residues between the homologue and the query in the window (i. e. N_{s}) with the length of the window (i. e. L_{w}). The residue with the highest P_{h} in the homologue is then privileged to assign its structure type (β-turn or non-β-turn) to the aligned residue in the query sequence if P_{h} meets either of the following conditions: (1)$P_{\mathrm{h}} \geqslant P_{\mathrm{q}}$; (2)$P_{\mathrm{h}} < P_{\mathrm{q}}$ and $P_{\mathrm{h}} > T_{\mathrm{l}}$, where P_{q} represents the probability value of NetTurnP prediction for the aligned residue in the query sequence, and T_l is a threshold value. If P_h meets neither conditions, the predicted type (β-turn or non-β-turn) by NetTurnP is retained for that residue.

1.5 Measures to evaluate our method

The quality of our method for classifying β-turn and non-β-turn is evaluated by five measures Q_{total}, Q_{pre}, Q_{obs} , MCC and AUC. Given that TP (true positives) is the number of correctly classified β-turn residues, TN (true negatives) is the number of correctly classified non-β-turn residues, FP (false positives) is the number of non-β-turn incorrectly classified as β-turn residues, and FN (false negatives) is the number of β-turn incorrectly classified as non-β-turn residues, Q_{total} (prediction accuracy) is defined as the percentage of correctly classified residues as

$$Q_{total}=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{4}$$

Q_{pre}, also called precision, is the percentage of correctly predicted β-turns among the predicted β-turns, i.e.

$$Q_{pre}=\frac{TP}{TP+FP}\times 100\% \tag{5}$$

Q_{obs}, also called sensitivity, is the percentage of correctly predicted β-turns among the observed (true) β-turns, i.e.

$$Q_{obs}=\frac{TP}{TP+FN}\times 100\% \tag{6}$$

Q_{total} is sometimes misleading, so a more robust measure is used, which is known as Matthews Correlation Coefficient (MCC).

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{7}$$

MCC can be in the range of −1 to 1, where 1 is a perfect correlation and −1 is the perfect anticorrelation. A value of 0 indicates no correlation. Higher MCC value corresponds to better performance of the prediction method. AUC, short for Area Under the Curve, is a threshold independent measure, and calculated from the receiver operating characteristic (ROC) curve which is a plot of the sensitivity against the False Positive rate = $FP/(FP+TN)$[49]. An AUC value above 0.7 is an indication of a useful prediction and a good prediction method achieves a value > 0.85 [50]. AUCCalculator is a Java jar file for calculating the AUC for both ROC graphs and Precision-Recall graphs [51], which was downloaded from http://mark.goadrich. com/programs/AUC/.

2 Results

2.1 Parameter optimization using EVA300

The EVA300 dataset was used to optimize the two parameters T_g and T_l of the structure mapping process. Given a fixed I_{max}, the grid search method was used to find the optimal combination of T_g and T_l according to the measure MCC. The search range of T_g is from 0.2 to 0.9 with a step of 0.05, and T_l changes from 0.5 to 0.9 with a step of 0.1 since $P_h \geqslant 0.5$ by definition and $P_h > T_l$, which require T_l to be at least 0.5 to take effect as a threshold. We varies the values of I_{max} itself from 0.2 to 1.0 with a step of 0.1 to investigate the performance variation of our method at different T_g, T_l combinations. Table 2 shows all optimal combinations found. Considering the robustness of different combinations at all I_{max} levels, we finally decided to use the optimal combination T_g=0.85, T_l=0.8 when I_{max}<0.5, and T_g=0.45, T_l=0.8 when I_{max}≥0.5.

Table 2 Parameter optimization using the EVA300 dataset

I_{max}	T_g	T_l	MCC
0.2	0.85	0.8	0.4704
0.3	0.85	0.8	0.4704
0.4	0.80	0.6	0.4725
0.4	0.85	0.8	0.4723
0.5	0.45	0.8	0.4809
0.6	0.45	0.8	0.4874
0.7	0.45	0.8	0.5005
0.8	0.45	0.8	0.5131
0.9	0.45	0.6	0.5273
0.9	0.45	0.8	0.5272
1.0	0.30	0.8	0.5425
1.0	0.45	0.8	0.5421

More decimal digits are retained for MCC than in **Table 1** to show its variation more precisely.

2.2 Evaluation of the method against BT426 and EVA937

Two datasets, BT426 and EVA937, were used to test the performance of our method for improving β-turn prediction. The results of our method were compared with those predicted by NetTurnP alone. For each query sequence of the two datasets, as mentioned in **Materials and methods**, the BLAST hits with sequence identity of more than 25% to the query were selected. We set the parameter I_{max} to 1.0 to fully employ homology information from the filtered PDB database. Before predicting on either of the BT426 and EVA937 datasets, the sequences in either dataset were removed from the PDB database. The selected hits must also be deposited earlier than the query sequence (controlled by the parameter "byDate" in our program), thus we can ensure that the query sequence itself was not selected as a hit. Table 3 shows the results of our method and NetTurnP on the two datasets. For BT426 dataset, the prediction accuracy of NetTurnP has been improved, *MCC* increases from 0.50 to 0.56 by 0.06, Q_{total} changes from 78.2% to 81.4% by 3.2% increase and the other three measures Q_{pre}, Q_{obs}, *AUC* also rise. For EVA937 dataset, the situation is similar, *MCC* increases from 0.46 to 0.52 by 0.06, Q_{total} from 77.3% to 80.4% by 3.1%, and the other three measures Q_{pre}, Q_{obs}, *AUC* also rise. ROC curves of the two methods on BT426 and EVA937 were plotted in Figure 1 and Figure 2, from which we can see the increase of *AUCs* more clearly.

Table 3 Comparsion between our method and NetTurnP against the BT426 and EVA937 datasets

Dataset	Predictor	Measure				
		MCC	Q_{total}/%	Q_{obs}/%	Q_{pre}/%	*AUC*
BT426	BTMapping	0.56	81.4	77.8	59.3	0.885
	NetTurnP	0.50	78.2	75.9	54.1	0.864
EVA937	BTMapping	0.52	80.4	72.6	59.4	0.858
	NetTurnP	0.46	77.3	70.6	54.1	0.838

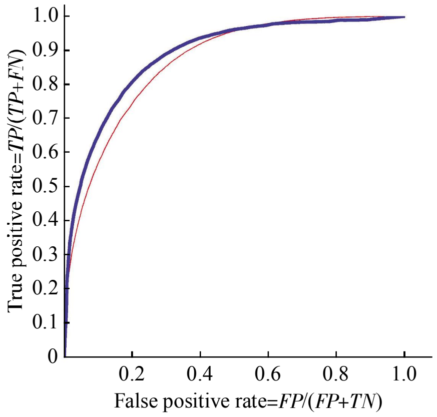

Fig. 1 ROC curves of NetTurnP and BTMapping on BT426 dataset

——: NetTurnP; ——: BTMapping.

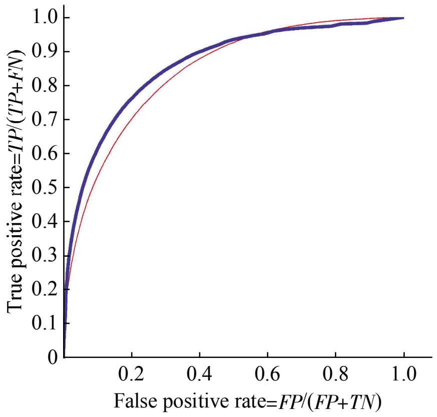

Fig. 2 ROC curves of NetTurnP and BTMapping on EVA937 dataset

——: NetTurnP; ——: BTMapping.

2.3 Evalution of the method by varying maximum sequence identity

In order to evaluate the performance of our method (i. e. BTMapping) under various conditions when the BLAST hits of the query sequences have different sequence identity levels, we vary $I_{\max}$ from 0.2 to 1.0 with a step of 0.1 to control the selection of the BLAST hits that are more than 100 residues in length. When $I_{\max}$ is small, only hits with low sequence identity are included for predicting, but when $I_{\max}$ gets bigger, hits with higher sequence identity are added. Both BT426 and EVA937 are used for evaluation. Table 4 lists all the results for the gradient $I_{\max}$ and the results of NetTurnP are also inserted for comparison, from which we can see that by employing the homology information from the BLAST hits, the accuracy of NetTurnP has been improved, and the most robust measure *MCC* increases as $I_{\max}$ increases for both datasets. Since the BLAST hits we selected must have more than 25% sequence identity to the query sequences as mentioned in **Materials and methods**, when $I_{\max}$ equals to 0.2, there are no BLAST hits selected with >100 residues, the main contribution to the increase of *MCC* comes from hits with ≤100 residues. When $I_{\max}$ equals to 0.3, all measures have the same values as those when $I_{\max}$ equals to 0.2, which means either the BLAST hits selected with > 0.25 and ≤0.3 sequence identity have little impact on increasing the *MCC* or there are no such hits found at all. When I_{max} equals to any value in the range of 0.4 to 1.0, as Table 4 shows, the BLAST hits selected with >100 residues start to obviously contribute to the increase of *MCC*.

Table 4 The performance of BTMapping at different maximum sequence identity(I_{max}) levels on BT426 and EVA937

Dataset	Predictor	I_{max}	Measure				
			MCC	Q_{total}/%	Q_{obs}/%	Q_{pre}/%	*AUC*
BT426	NetTurnP	–	0.4972	78.24	75.85	54.11	0.864
	BTMapping	0.2	0.5013	78.45	76.07	54.42	0.866
		0.3	0.5013	78.45	76.07	54.42	0.866
		0.4	0.5020	78.49	76.06	54.50	0.866
		0.5	0.5085	78.92	76.03	55.19	0.868
		0.6	0.5198	79.53	76.40	56.16	0.871
		0.7	0.5260	79.89	76.52	56.76	0.872
		0.8	0.5276	79.97	76.58	56.89	0.873
		0.9	0.5319	80.19	76.74	57.26	0.874
		1.0	0.5561	81.39	77.84	59.27	0.885
EVA937	NetTurnP	–	0.4638	77.28	70.59	54.09	0.838
	BTMapping	0.2	0.4700	77.61	70.82	54.60	0.840
		0.3	0.4700	77.61	70.82	54.60	0.840
		0.4	0.4706	77.66	70.78	54.69	0.840
		0.5	0.4790	78.17	70.82	55.55	0.842
		0.6	0.4868	78.60	70.98	56.28	0.845
		0.7	0.4952	79.04	71.22	57.04	0.848
		0.8	0.4983	79.21	71.30	57.34	0.848
		0.9	0.5009	79.33	71.45	57.54	0.849
		1.0	0.5234	80.40	72.56	59.37	0.858

More decimal digits are retained for the first four measures than in Table 1 to illustrate the change of them more clearly.

2.4 The impact of "byDate" value on prediction accuracy and comparison with the predictor ShapeString_Pred

The parameter "byDate" in our program controls the selection of BLAST hits, only the hits deposited earlier than the date represented by "byDate" are adopted for structure mapping. Therefore, set with different "byDate" values, our method may have different prediction accuracies. For a newly identified protein sequence, only structures deposited earlier in the PDB can actually be employed for structure mapping. BT426 was used for this testing. As mentioned in **Materials and methods**, the recent release of PDB sequences was filtered by CD-HIT[48] utility at 95% sequence identity threshold and the included BT426 sequences were removed from it. We varied "byDate" from 1st January 1990 to 1st January 2011 with a step of one year. Additionally, the date 1st September 2010 was also considered for comparison with ShapeString_Pred. Table 5 reports the result. It can be concluded from the table that as the "byDate" value increases, *MCC* and Q_{total} both increase, which means the prediction accuracy becomes higher. We believe the reason for this is that as the "byDate" value gets nearer, more homologues to a query sequence can be found through BLAST, and it is of higher possibility to find optimal homologues for structure mapping to increase the prediction accuracy. In order to compare with the ShapeString_Pred, we constructed a release of PDB sequences by applying a 30% sequence identity cutting using CD-HIT[48] and removed the included BT426 sequences as what was done in the evaluation of ShapeString_Pred. Since the release of PDB sequences adopted by ShapeString_Pred was downloaded in September, 2010, we set the "byDate" parameter to 01-SEP-10. As Table 5 shows, we achieved a result of: *MCC*=0.7152, Q_{total}=89.0%, Q_{obs}= 83.2%, Q_{pre}=74.8%, and this is better than the 7-fold cross-validation result of ShapeString_Pred: *MCC* = 0.66, Q_{total}=87.2%, Q_{obs}=75.9%, Q_{pre}=73.8%.

Table 5 Performance of the method on BT426 dataset at different "byDate" values

byDate	Measure				
	MCC	Q_{total}/%	Q_{obs}/%	Q_{pre}/%	*AUC*
01-Jan-90	0.5047	78.7	76.1	54.7	0.866
01-Jan-91	0.5064	78.7	76.2	54.9	0.867
01-Jan-92	0.5101	78.9	76.4	55.2	0.869
01-Jan-93	0.5202	79.5	76.6	56.1	0.873
01-Jan-94	0.5352	80.3	77.0	57.5	0.879
01-Jan-95	0.5517	81.2	77.6	58.9	0.884
01-Jan-96	0.5773	82.5	78.4	61.3	0.893
01-Jan-97	0.5955	83.4	79.1	62.9	0.898
01-Jan-98	0.6300	85.0	80.5	66.1	0.909
01-Jan-99	0.6632	86.5	82.1	69.0	0.918
01-Jan-00	0.6972	88.1	83.4	72.3	0.926
01-Jan-01	0.7245	89.3	84.5	75.0	0.931
01-Jan-02	0.7321	89.6	84.9	75.7	0.933
01-Jan-03	0.7416	90.0	85.1	76.7	0.935
01-Jan-04	0.7503	90.4	85.4	77.7	0.935
01-Jan-05	0.7540	90.5	85.8	77.8	0.936
01-Jan-06	0.7611	90.8	86.0	78.6	0.937
01-Jan-07	0.7650	91.0	86.1	79.0	0.937
01-Jan-08	0.7747	91.4	86.5	80.0	0.938
01-Jan-09	0.7805	91.6	86.9	80.5	0.938
01-Jan-10	0.7877	91.9	87.2	81.3	0.939
01-Jan-11	0.7897	92.0	87.3	81.4	0.939
[a]01-Sep-10	0.7898	92.0	87.4	81.4	0.939
[b]01-Sep-10	0.7152	89.0	83.2	74.8	0.921

a. The release of PDB sequences was filtered by CD-HIT at 95% sequence identity threshold. b. The release of PDB sequences was filtered by CD-HIT at 30% sequence identity threshold, same as what was done in the paper introducing the two-layer SVM predictor ShapeString_Pred[38]. More decimal digits are retained for the measure *MCC* than in Table 1 to illustrate its variation more precisely.

3 Discussion

Testing on BT426 and EVA937, by using the structures of homologues deposited earlier in PDB than the query sequences to be predicted, we can improve the prediction accuracy of a neural network β-turn predictor NetTurnP. In order to evaluate the performance of our method (i. e. BTMapping) under various conditions when the BLAST hits have different sequence identity levels, we varied the I_{max} to simulate such situations, and we found that even when I_{max} is pretty low, the prediction accuracy of NetTurnP can be improved. We believe this is contributed by the short homologues (≤100 residues) and part of the long homologues (>100 residues) with low I_{max}. As the structural databases like PDB get larger, the possibility of having homologues for a newly identified sequence is higher and it is easier to be predicted accurately. We varied the "byDate" parameter to see the performance of our method, we found that when the "byDate" value is closer to the present time, the prediction accuracy is higher, which means more or better homologues are found. Under the same conditions, comparing with the recent two-layer SVM predictor ShapeString_Pred that uses homology information from PDB in a different way, our method shows better performance. Besides NetTurnP, our method can be easily integrated with other de novo β-turn predictors to improve their performance as long as their outputs of prediction are formatted as required.

There is still room for improvement in that it can only conduct two-class (β-turn or non-β-turn) prediction currently. Since there are many β-turn subtypes, we hope to extend our method for predicting β-turn subtypes in the future work. The algorithm of the mapping process in our work is relatively simple, which would allow integration with more sophisticated algorithms to further improve the β-turn prediction accuracy.

Acknowledgements The authors would like to thank Dr. Mingshun Chen at USDA-ARS and Department of Entomology, Kansas State University, Manhattan, Kansas, USA for his comments and suggestions during the manuscript preparation.

References

[1] Garg A, Kaur H, Raghava G P. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins, 2005, **61** (2): 318-324

[2] Chen K, Kurgan L. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics, 2007, **23** (21): 2843-2850

[3] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA, 2004, **101** (24): 8942-8944

[4] Fuchs P F J, Alix A J P. High accuracy prediction of β-turns and their types using propensities and multiple alignments. Proteins: Structure, Function, and Bioinformatics, 2005, **59** (4): 828-839

[5] Wang Y, Xue Z, Xu J. Better prediction of the location of alpha-turns in proteins with support vector machine. Proteins, 2006, **65** (1): 49-54

[6] Song J, Burrage K. Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinformatics, 2006, **7**: 425

[7] Kim D E, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res, 2004, **32** (Web Server issue): W526-531

[8] McGuffin L J, Bryson K, Jones D T. The PSIPRED protein structure prediction server. Bioinformatics, 2000, **16** (4): 404-405

[9] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983, **22** (12): 2577-2637

[10] Richardson J S. The anatomy and taxonomy of protein structure. Adv Protein Chem, 1981, **34**: 167-339

[11] Takano K, Yamagata Y, Yutani K. Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. Biochemistry, 2000, **39** (29): 8655-8665

[12] Trevino S R, Schaefer S, Scholtz J M, *et al*. Increasing protein conformational stability by optimizing beta-turn sequence. J Mol Biol, 2007, **373** (1): 211-218

[13] Fu H, Grimsley G R, Razvi A, *et al*. Increasing protein stability by improving beta-turns. Proteins, 2009, **77** (3): 491-498

[14] de la Cruz X, Hutchinson E G, Shepherd A, *et al*. Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci USA, 2002, **99** (17): 11157-11162

[15] Kumar M, Bhasin M, Natt N K, *et al*. BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res, 2005, **33** (Web Server issue): W154-159

[16] Rose G D, Gierasch L M, Smith J A. Turns in peptides and proteins. Adv Protein Chem, 1985, **37**: 1-109

[17] Muller G, Hessler G, Decornez H Y. Are beta-turn mimetics mimics of beta-turns?. Angew Chem Int Ed Engl, 2000, **39** (5): 894-896

[18] Kee K S, Jois S D. Design of beta-turn based therapeutic agents. Curr Pharm Des, 2003, **9** (15): 1209-1224

[19] Fuller A A, Du D, Liu F, *et al*. Evaluating beta-turn mimics as beta-sheet folding nucleators. Proc Natl Acad Sci USA, 2009, **106** (27): 11067-11072

[20] Wilmot C M, Thornton J M. Beta-turns and their distortions: a proposed new nomenclature. Protein Eng, 1990, **3** (6): 479-493

[21] Zhang C-T, Chou K-C. Prediction of β-turns in proteins by 1-4 and 2-3 correlation model. Biopolymers, 1997, **41** (6): 673-702

[22] Chou P Y, Fasman G D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry, 1974, **13** (2): 211-222

[23] Wilmot C M, Thornton J M. Analysis and prediction of the different types of beta-turn in proteins. J Mol Biol, 1988, **203** (1): 221-232

[24] Chou K C, Blinn J R. Classification and prediction of beta-turn types. J Protein Chem, 1997, **16** (6): 575-595

[25] McGregor M J, Flores T P, Sternberg M J. Prediction of beta-turns in proteins using neural networks. Protein Eng, 1989, **2** (7): 521-526

[26] Shepherd A J, Gorse D, Thornton J M. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci, 1999, **8** (5): 1045-1055

[27] Kaur H, Raghava G P. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci, 2003, **12** (3): 627-634

[28] Kirschner A, Frishman D. Prediction of β-turns and β-turn types by a novel bidirectional Elman-type recurrent neural network with multiple output layers (MOLEBRNN). Gene, 2008, **422** (1-2): 22-29

[29] Kaur H, Raghava G P. A neural network method for prediction of beta-turn types in proteins using evolutionary information. Bioinformatics, 2004, **20** (16): 2751-2758

[30] Petersen B, Lundegaard C, Petersen T N. NetTurnP--neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. PLoS One, 2010, **5** (11): e15079

[31] Kim S. Protein beta-turn prediction using nearest-neighbor method. Bioinformatics, 2004, **20** (1): 40-44

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

[33] Zhang Q, Yoon S, Welsh W J. Improved method for predicting beta-turn using support vector machine. Bioinformatics, 2005, **21** (10): 2370-2374

[34] Zheng C, Kurgan L. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics, 2008, **9** (1): 430

[35] Hu X, Li Q. Using support vector machine to predict β- and γ-turns in proteins. J Computational Chemistry, 2008, **29** (12): 1867-1875

[36] Liu L, Fang Y, Li M, *et al*. Prediction of Beta-turn in protein using E-sspred and support vector machine. Protein J, 2009, **28** (3-4): 175-181

[37] Kountouris P, Hirst J D. Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures. BMC Bioinformatics, 2010, **11** (1): 407

[38] Tang Z, Li T, Liu R, *et al*. Improving the performance of beta-turn prediction using predicted shape strings and a two-layer support vector machine model. BMC Bioinformatics, 2011, **12** (1): 283

[39] Guruprasad K, Rajkumar S. Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. J Biosci, 2000, **25** (2): 143-156

[40] Montgomerie S, Sundararaj S, Gallin W J, *et al*. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics, 2006, **7**: 301

[41] Amegbey G, Stothard P, Kuznetsova E, *et al*. Solution structure of MTH0776 from methanobacterium thermoautotrophicum. J Biomol NMR, 2005, **33** (1): 51-56

[42] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool. J Mol Biol, 1990, **215** (3): 403-410

[43] Chenna R, Sugawara H, Koike T, *et al*. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res, 2003, **31** (13): 3497-3500

[44] Larkin M A, Blackshields G, Brown N P, *et al*. Clustal W and Clustal X version 2.0. Bioinformatics, 2007, **23** (21): 2947-2948

[45] Kaur H, Raghava G P. An evaluation of beta-turn prediction methods. Bioinformatics, 2002, **18** (11): 1508-1514

[46] Hutchinson E G, Thornton J M. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci, 1996, **5** (2): 212-220

[47] Koh I Y, Eyrich V A, Marti-Renom M A, *et al*. EVA: Evaluation of protein structure prediction servers. Nucleic Acids Res, 2003, **31** (13): 3311-3315

[48] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 2001, **17** (3): 282-283

[49] Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Netherlands: Kluwer Academic Publish, 2004

[50] Lund O, Nielsen M, Lundegaard C, *et al*. Immunological Bioinformatics (Computational Molecular Biology). USA: The MIT Press, 2005

[51] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; Pittsburgh, Pennsylvania. 1143874: ACM, 2006

运用 PDB 中的同源信息提高 NetTurnP 的蛋白质 β 转角预测精度 *

钱 刚 [1, 2)] 王海燕 [3)] 袁哲明 [1, 2)**]

([1)] 湖南省作物种质创新与资源利用重点实验室，长沙 410128；[2)] 湖南省植物病虫害生物学与防控重点实验室，长沙 410128；
[3)] Department of Statistics, Kansas State University, Manhattan, Kansas 66506, USA)

摘要 β 转角作为一种蛋白质二级结构类型在蛋白质折叠、蛋白质稳定性、分子识别等方面具有重要作用．现有的 β 转角预测方法，没有将 PDB 等结构数据库中先前存在的同源序列的结构信息映射到待预测的蛋白质序列上．PDB 存储的结构已超过 70 000，因此对一条新确定的序列，有较大可能性从 PDB 中找到其同源序列．本文融合 PDB 中提取的同源结构信息(对每一待测序列，仅使用先于该序列存储于 PDB 中的同源信息)与 NetTurnP 预测，提出了一种新的 β 转角预测方法 BTMapping，在经典的 BT426 数据集和本文构建的数据集 EVA937 上，以马修斯相关系数表示的预测精度分别为 0.56、0.52，而仅使用 NetTurnP 的为 0.50、0.46，以 Q_{total} 表示的预测精度分别为 81.4%、 80.4%，而仅使用 NetTurnP 的为 78.2%、77.3%．结果证实同源结构信息结合先进的 β 转角预测器如 NetTurnP 有助于改进 β 转角识别．BTMapping 程序及相关数据集可从 http:// www.bio530.weebly.com 获得．

关键词 β 转角预测，同源信息，PDB，NetTurnP，BTMapping

学科分类号 Q518.1 **DOI**: 10.3724/SP.J.1206.2011.00370

* 湖南省杰出青年科学基金(10JJ1005)和高等学校博士点基金(200805370002)资助项目.

** 通讯联系人.

Tel: 0731-84613956, E-mail: zhmyuan@sina.com

收稿日期：2011-08-11，接受日期：2011-11-04