

# Prediction and Analysis of $\beta$ -Turns in Proteins by Support Vector Machine

Tho Hoan Pham<sup>1</sup>    Kenji Satou<sup>1,2</sup>    Tu Bao Ho<sup>1</sup>  
h-pham@jaist.ac.jp    ken@jaist.ac.jp    bao@jaist.ac.jp

<sup>1</sup> Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

<sup>2</sup> Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Corporation (JST)

## Abstract

Tight turn has long been recognized as one of the three important features of proteins after the  $\alpha$ -helix and  $\beta$ -sheet. Tight turns play an important role in globular proteins from both the structural and functional points of view. More than 90% tight turns are  $\beta$ -turns. Analysis and prediction of  $\beta$ -turns in particular and tight turns in general are very useful for the design of new molecules such as drugs, pesticides, and antigens. In this paper, we introduce a support vector machine (SVM) approach to prediction and analysis of  $\beta$ -turns. We have investigated two aspects of applying SVM to the prediction and analysis of  $\beta$ -turns. First, we developed a new SVM method, called BTSVM, which predicts  $\beta$ -turns of a protein from its sequence. The prediction results on the dataset of 426 non-homologous protein chains by sevenfold cross-validation technique showed that our method is superior to the other previous methods. Second, we analyzed how amino acid positions support (or prevent) the formation of  $\beta$ -turns based on the “multivariable” classification model of a linear SVM. This model is more general than the other ones of previous statistical methods. Our analysis results are more comprehensive and easier to use than previously published analysis results.

**Keywords:**  $\beta$ -turns, protein secondary structure, support vector machine, support of amino acid position

**Availability:** The software is available based on requests at: <http://www.jaist.ac.jp/~h-pham/btsvm/>.

**Contact:** h-pham@jaist.ac.jp.

## 1 Introduction

Tight turns [18] play an important role in protein folding and stability. Tight turns are classified as  $\sigma$ -turns,  $\gamma$ -turns,  $\beta$ -turns,  $\alpha$ -turns, and  $\pi$ -turns.  $\beta$ -turn is a four-residue reversal in a protein chain that is not in an  $\alpha$ -helix, and the distance between  $C_{\alpha}(i)$  and  $C_{\alpha}(i+3)$  is lesser than 7Å [15, 16]. The  $\beta$ -turns are the most commonly found turns and make up about one-fourth of all residues in proteins.  $\beta$ -turns provide very useful information for defining template structures for the design of new molecules such as drugs, pesticides, and antigens.

There have been some attempts for prediction and analysis of  $\beta$ -turns. They can be divided into two categories: statistics-based and machine learning-based methods. The majority of statistics-based methods empirically employed the “positional preference approaches” [4, 5, 20, 21]. A machine learning-based method, called BTPRED, has been recently developed for prediction of  $\beta$ -turns and significantly outperformed statistics-based methods [12, 17]. The prediction accuracy of BTPRED is about  $Q_{total} = 71.6$ ,  $Q_{pred} = 44.1$ ,  $Q_{obs} = 57.3$  and  $MCC = 0.31$  when using only single sequence and  $Q_{total} = 73.5$ ,  $Q_{pred} = 47.2$ ,  $Q_{obs} = 64.3$  and  $MCC = 0.37$  when using multiple sequence alignment

( $Q_{total}$ ,  $Q_{pred}$ ,  $Q_{obs}$  and  $MCC$  are defined later in the text). The accuracy has been further improved to  $Q_{total} = 75.5$ ,  $Q_{pred} = 49.8$ ,  $Q_{obs} = 72.3$  and  $MCC = 0.43$  by a couple of neural networks and the additional secondary structure information.

In this paper, we introduce another machine learning method, support vector machine (SVM), for both prediction and analysis of  $\beta$ -turns. SVM is based on statistical learning theory and was developed for the first time by Vapnik [7, 19]. In practice, SVM has good performance and is easier to implement and train than neural networks. It is a promising technique that has been successfully applied to problems in bioinformatics such as secondary structure prediction [13], microarray data analysis [9], protein-protein interactions [14], fold recognition, translation initiation site recognition, etc.

Two aspects of applying SVM to prediction and analysis of  $\beta$ -turns have been investigated in this research. First, we develop a method, called BTSVM, that predicts  $\beta$ -turns of a protein from its sequence using SVM. The prediction can be done with single sequence or multiple sequence alignment. Our prediction results on the dataset of 426 non-homologous protein chains by sevenfold cross-validation technique showed that BTSVM performed very well when compared to BTPRED and the other methods. The accuracy of BTSVM is  $Q_{total} = 74.2$ ,  $Q_{pred} = 47.6$ ,  $Q_{obs} = 49.2$  and  $MCC = 0.31$  with single sequence, and  $Q_{total} = 78.4$ ,  $Q_{pred} = 55.9$ ,  $Q_{obs} = 58.6$  and  $MCC = 0.43$  with multiple alignment (PSSM). Furthermore, the result of our method further improved to  $Q_{total} = 78.7$ ,  $Q_{pred} = 56.0$ ,  $Q_{obs} = 62.0$  and  $MCC = 0.45$  when combined with the additional secondary structure information, which is in turn predicted by another high accuracy secondary structure prediction method such as PSIPRED. Moreover, by indicating specifically whether four consecutive residues in a protein form a  $\beta$ -turn or not at the same time, our method gives prediction results more clearly than BTPRED.

Second, we analyzed  $\beta$ -turns by proposing a new concept of “*the support of an amino acid position to the formation of  $\beta$ -turns under a linear SVM classification model*” (we refer to it as *the support of an amino acid position* for short), which implies both the contribution and prevention of that amino acid to the formation of  $\beta$ -turns. This information can be easily extracted from the “multivariable” classification model of a trained linear SVM. This model is more general than previously proposed models for prediction and analysis of  $\beta$ -turns such as Site-Independent model [6], 1-4 and 2-3 Residue-Correlation model [21], and Sequence-Couple model [4]. Our analysis results, based on the supports of amino acid positions, are more comprehensive and easier to use than the previous ones.

Our method for predicting  $\beta$ -turns with high accuracy and our easily understood analysis results will be helpful for the researchers working in the fields of fold recognition and design of new molecules.

## 2 Materials and Methods

### 2.1 The Dataset

To compare our method with the previous ones, we selected the dataset of 426 non-homologous protein chains described in the work of Guruprasad and Rajkumar [8]. This dataset has been used by Kaur and Raghava [11, 12] for assessing the performance of the  $\beta$ -turn prediction methods. In this dataset, there are no two protein chains having more than 25% sequence identity. The structure of these proteins is determined by X-ray crystallography at more than 2.0 Å resolution. Each chain contains at least one  $\beta$ -turn. The program PROMOTIF [10] has been used to assign  $\beta$ -turns in proteins.

### 2.2 Vector Representations of a Protein Sequence

There are two basic ways to represent a protein sequence.

1. Single sequence: Each residue is represented by a 20-dimension vector of 0 and 1 that is a coding for the corresponding amino acid at this residue. This binary representation can be extended by taking into account the general substitute abilities (scores) of amino acids, i.e. BLOSUM62.

Therefore, each residue is represented by a 20-dimension vector of the substitute scores of 20 amino acids for this residue.

- Multiple sequence alignment: A protein sequence is firstly aligned with a non-redundant (NR) database (e.g., the version used in our work contains 1,109,366 sequences) to find the family of sequences to which that protein belongs. The alignment can be expressed in a scoring matrix of probability estimates or scores [1]. Two kinds of such matrices are considered in our work: position-specific frequent matrices (PSFM) and position-specific scoring matrices (PSSM). PSFM and PSSM in this work are in PSI-BLAST profiles with E-value threshold of 0.001 and three iterations.

In these two ways, each protein sequence is represented as a bi-dimensional vector  $L \times 20$ , where  $L$  is the length of the sequence. In this work, all elements of bi-dimensional vectors are scaled into the interval  $[-1, 1]$  by a simple linear transformation function before moving into the machine learning system (i.e. support vector machine).

### 2.3 SVM Method for Prediction of $\beta$ -Turns

Support Vector Machine (SVM) is a technique of machine learning based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, map the input vectors into a feature space (often with a higher dimension), either linearly or non-linearly, which is relevant with the selection of the kernel function. Second, seek an optimized linear division within the feature space from the first step, i.e. construct a hyperplane which separates two classes. SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description of the theory of SVMs for pattern recognition has been done by Vapnik [19].

In this paper, we apply Vapnik's Support Vector Machine for prediction and analysis of  $\beta$ -turns in proteins. Our method, BTSVM, used the SVM libraries in the software LIBSVM [2], which is an implementation of SVM in C language for the problem of classification.

To predict  $\beta$ -turns, we used sliding windows of size  $w \geq 4$  along the vector representation of a protein. Therefore, each window is a bi-dimension vector  $w \times 20$ . The prediction based on SVM (like other machine learning techniques) includes two phases: training and testing. In the training phase, a window (of the size  $w$ ) with four central residues forming a  $\beta$ -turn is considered as a positive sample (called *turn-window*), otherwise as a negative sample (called *non-turn-window*). For example, the following protein sequence, with states of T, t for  $\beta$ -turn (T means the beginning state of a  $\beta$ -turn) and n for non- $\beta$ -turn in the next line, contains 4 turn-windows and 16 non-turn-windows if we use sliding windows of length  $w = 8$ .

```
GSVGGMLGLPFNDVYCASKFALEGLCE
nnnnnnnnTTtttnnnTtTtttnnn
```

We assume that  $\{X_i, y_i\}_{i=1, \dots, l}$  be a training set, where  $X_i$  is a bi-dimensional vector  $w \times 20$  of scores (when using PSSM) or frequencies (when using PSFM), and  $y_i \in \{0, 1\}$ . SVM solves the following primal problem:

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \\ & y_i (w^T \phi(X_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \dots, l \end{aligned}$$

Its dual is a quadratic optimization problem:

$$\min_{\alpha} \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

$$0 \leq \alpha_i \leq C$$

$$y^T \alpha = 0$$

where  $e$  is the vector of all ones;  $C > 0$  is a error penalty parameter;  $y = \{y_i\}_{i=1,\dots,l}$ ;  $Q_{ij} = y_i y_j K(X_i, X_j)$ ;  $K(X_i, X_j) = \phi(X_i)^T \phi(X_j)$  is a kernel function; and  $\phi(X_i)$  maps  $X_i$  into a higher (maybe infinite) dimensional space. So  $K(X_i, X_j)$  is a positive definite function that reflects the similarity between the sample  $X_i$  and the sample  $X_j$ . In this work, we employed linear functions ( $K(X_i, X_j) = \langle X_i, X_j \rangle$ ) and radial basis functions ( $K(X_i, X_j) = \exp(-\gamma(X_i - X_j)^2)$ ) as the kernel functions. The SVM classification function has the following form:

$$f(X) = \sum_i \alpha_i y_i K(X, X_i) + b \quad (1)$$

where  $\alpha = \{\alpha_i\}_{i=1,\dots,l}$  is the solution of the above dual problem and  $b$  is in the solution of the primal problem. In Equation 1, vectors  $\{X_i\}$  that correspond with  $\alpha_i > 0$  are called as ‘‘support vectors’’.

In the testing phase, conversely, for each sliding window  $X$ , if  $f(X) > 0$ ,  $X$  is predicted as a turn-window (or non-turn-window); otherwise  $X$  is predicted as a non-turn-window (or turn-window). This prediction result is assigned to four residues in the middle of  $X$  (if  $X$  is a turn-window, the prediction for four central residues is ‘‘Tttt’’). Therefore, in a given protein, each residue is assigned four times when the window slides along the sequence. The final prediction result assigned to a residue is  $\beta$ -turn if at least one of four assignments for this residue is a  $\beta$ -turn.

## 2.4 Support of an Amino Acid Position to the Formation of $\beta$ -Turns

The identification of discriminant amino acid positions is helpful not only in predicting  $\beta$ -turns but also in designing turns in proteins [20, 8, 3]. Each amino acid position, in our situation, is represented by an amino acid positional score (when using PSSM) or amino acid positional frequency (when using PSFM). There is a fact that some amino acid positions in a window contribute to the formation of  $\beta$ -turn (turn-window, i.e. the  $\beta$ -turn formation of four central residues in the window), while some others, on the other hand, prevent the formation of  $\beta$ -turn. The contribution (or prevention) of an amino acid position to the formation of  $\beta$ -turns can be assessed by the support of that amino acid position in the classification function of a trained SVM (Equation 1). The best classification function may (even usually) be the nonlinear function, but it is difficult to assess the support of a particular amino acid position to the classification function itself. Therefore, in this work, we used BTSVM with a linear kernel. The classification function of the linear BTSVM (Equation 1) then has the following form:

$$f(X) = \sum_{a \in \text{amino\_acids}, i \in \{1,2,\dots,w\}} \beta_{ai} X_{ai} + \rho \quad (2)$$

where  $X = (X_{ai})$  is a sliding window of scores (frequencies) of amino acid positions from PSSM (PSFM);  $X_{ai}$  is score (frequency) of amino acid  $a$  at position  $i$  in the multiple sequence alignment. We can exchange the sign of the weights  $\beta_{ai}$  and  $\rho$  so that if  $f(X) > 0$  then  $X$  is classified as a  $\beta$ -turn (i.e. having four central residues form a  $\beta$ -turn), else  $X$  is classified as a non- $\beta$ -turn. This classification rule is called the classification model of the linear SVM.

We define the weight  $\beta_{ai}$  as *the support of amino acid position ai (amino acid a at position i) for the formation of  $\beta$ -turns under a linear SVM classification model (or the support of amino acid position for short)*. If this support  $\beta_{ai}$  is positive (*positive support*), the amino acid position  $ai$  would contribute to the formation of  $\beta$ -turns; if it is negative (*negative support*), the amino acid position  $ai$  would prevent the formation of  $\beta$ -turns; and the larger the absolute value of this support, the stronger the contribution (or prevention if this support is negative).

## 2.5 Performance Measures

In this research, we used four criteria as described in the paper of Shepherd et al. [17]: (1)  $Q_{total}$  (or prediction accuracy), the percentage of correctly predicted residues, (2) Matthew's Correlation Coefficient (MCC) accounts for both over- and under-prediction, (3)  $Q_{pred}$ , the percentage of correct prediction of  $\beta$ -turns (or probability of correct prediction), and (4)  $Q_{obs}$ , the percentage of observed  $\beta$ -turns that are correctly predicted (or percent coverage). The parameters can be calculated by the following equations:

$$Q_{total} = \left(\frac{p+n}{t}\right) \times 100$$

$$MCC = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}$$

$$Q_{pred} = \left(\frac{p}{p+o}\right) \times 100$$

$$Q_{obs} = \left(\frac{p}{p+u}\right) \times 100$$

where  $p$  and  $n$  are number of correctly predicted  $\beta$ -turn and non- $\beta$ -turn residues, respectively;  $o$  and  $u$  are the number of incorrectly predicted  $\beta$ -turn and non- $\beta$ -turn residues, respectively; and  $t = p + n + o + u$  is the total of residues.

## 2.6 Sevenfold Cross-Validation

For comparing our method to the other ones, we employed the sevenfold cross-validation described in the work of Kaur and Raghava [11]. The dataset of 426 non-homologous protein chains is randomly divided into seven subsets, each containing equal number of proteins. Each set is an unbalanced set that retains the naturally occurring proportion of  $\beta$ -turns and non-turns. Five of seven subsets are grouped into the training set. The SVM is validated for minimum error on the sixth subset (validation set) to avoid over-training. The last subset is for the testing set. This has been done seven times to test the prediction result for each testing set. The final prediction results have been averaged over seven testing sets.

## 3 Results

### 3.1 Prediction of $\beta$ -Turns

In this work, we used BTSVM with a radial basis function kernel ( $\gamma = 0.01$ ), PSSM and sliding windows of length 12. These were selected in order to get the maximum performance from various experiments on the validation sets. As can be seen in Table 1, the performance of our method, BTSVM, compares well to the others on the dataset of 426 non-homologous protein chains by sevenfold cross-validation. Our method gave the best performance of  $Q_{total} = 78.4$ ,  $Q_{pred} = 55.9$ ,  $Q_{obs} = 58.6$  and  $MCC = 0.43$  when using multiple sequence alignments. The values of the two most important performance parameters,  $Q_{total}$  and  $MCC$ , obtained by our method are 4.9 and 0.06 respectively greater than those obtained with BTPRED, which has employed neural network techniques, and much greater than values obtained with statistical methods.

As in the work of Kaur and Raghava (2003), although we know that it may be unfair to use the additional secondary structure information, which is directly predicted by the PSIPRED method without re-training it in the training dataset, we tried to combine our prediction results with the secondary structure information in a simple way. First, we appropriately controlled the bias for  $\beta$ -turn prediction in our method and then removed (changed to non-turn) those predicted  $\beta$ -turn windows that contained more than two residues (of the four central residues) falling in  $\alpha$ -helix or

Table 1: Results of  $\beta$ -turn/non- $\beta$ -turn predictions. The results of Chou-Fasman, Thornton, 1-4 & 2-3 correlation model and sequence couple model at original and new (in brackets) threshold values are from [11]. The results of BTPRED are from [12]. The results of BTSVM are sevenfold cross-validation accuracies obtained in the same way. BTSVM-LIN is used for analysis of  $\beta$ -turns. BTSVM1 is a SVM method that predicts  $\beta$ -turns on only the single central residue of a sliding window.

		$Q_{total}$	$Q_{pred}$	$Q_{obs}$	$MCC$
Chou-Fasman	Sin. seq.	74.9 (69.3)	46.1 (36.9)	16.9 (35.3)	0.16 (0.16)
	Sin. seq. & sec. struct.	74.3 (75.3)	47.7 (49.6)	54.3 (47.5)	0.34 (0.32)
Thornton	Sin. seq.	74.5 (70.1)	44.0 (36.7)	16.7 (30.5)	0.15 (0.14)
	Sin. seq. & sec. struct.	75.2 (75.2)	49.3 (49.3)	44.9 (44.9)	0.31 (0.31)
1-4 & 2-3 correlation model	Sin. seq.	63.2 (71.1)	35.3 (40.8)	60.4 (40.3)	0.21 (0.21)
	Sin. seq. & sec. struct.	73.4 (74.8)	46.2 (48.0)	51.5 (39.8)	0.31 (0.28)
Sequence couple model	Sin. seq.	50.6 (72.7)	31.7 (43.9)	88.4 (41.0)	0.23 (0.25)
	Sin. seq. & sec. struct.	72.2 (75.4)	45.0 (49.6)	60.0 (40.0)	0.33 (0.28)
BTPRED	Sin. seq.	71.6	44.1	57.3	0.31
	Mul. seq.	73.5	47.2	64.3	0.37
	Mul. seq. & sec. struct.	75.5	49.8	72.3	0.43
BTSVM	Sin. seq.	74.2	47.6	49.2	0.31
	Mul. seq.	78.4	55.9	58.6	0.43
	Mul. seq. & sec. struct.	78.7	56.0	62.0	0.45
BTSVM-LIN	Mul. seq. (PSFM)	73.1	46.0	55.0	0.32
BTSVM1	Mul. seq. (PSSM)	75.8	50.7	70.7	0.437

$\beta$ -strand regions in PSIPRED prediction results (see Table 2). The performance is improved to  $Q_{total} = 78.7$ ,  $Q_{pred} = 56.0$ ,  $Q_{obs} = 62.0$  and  $MCC = 0.45$ .

We tried to apply SVM to predict  $\beta$ -turns on only the single central residue (BTSVM1 in the Table 1). Although the performance of  $MCC$  slightly increases when compared to prediction on the four central residues simultaneously (BTSVM), the total prediction accuracy,  $Q_{total}$ , decreases from 78.4 to 75.8. Moreover, the prediction results of BTSVM1 are not only invalid in the cases that less than four consecutive residues are predicted as “turn” and but also unclear in the cases that more than five consecutive residues are predicted as “turn”.

### 3.2 Supports of Amino Acid Positions to the Formation of $\beta$ -Turns

In this work, we used PSFM instead of PSSM (since PSSM incorporates the abilities of amino acid replacement, it is not suitable for this work), BTSVM with a linear kernel (BTSVM-Lin) (see materials and methods), sliding windows with a length of 8 (this length was selected in order to get the maximum performance). After training on the dataset of 426 non-homologous protein chains, the classification model of BTSVM-Lin is linear and its classification function has the form of Equation 2. Table 3 presents the supports,  $\alpha_{ai}$ , of the amino acid positions to the formation of  $\beta$ -turns (turn-windows) under this linear classification model.

As can be seen, the supports of the amino acid positions are like a “spectrum” from the smallest value, -1.250 (amino acid Ile at position 3), to the largest value, 2.712 (amino acid Asn at position 5). Amino acid positions which most strongly support the formation of  $\beta$ -turns are shown in Table 2 by boldface, and those which weakest (negative) support the formation of  $\beta$ -turns are underlined.

In general, the supports of amino acids for the formation of  $\beta$ -turns are different and vary from position to position. Amino acids Ser (S) and Cys (C) nearly neither contribute nor prevent the

Table 2: Dependence of prediction results on the error penalty cost of  $\beta$ -turn prediction. (\*): the best performances.

		$Q_{total}$	$Q_{pred}$	$Q_{obs}$	$MCC$
$C_{turn} = 3.5, C_{nonturn} = 1$	BTSVM	79.7	62.5	44.1	0.40
	BTSVM + PSIPRED	80.2	64.7	43.1	0.41
$C_{turn} = 4.0, C_{nonturn} = 1$	BTSVM	79.3	58.9	52.8	0.42
	BTSVM + PSIPRED	80.1	61.3	51.5	0.44
$C_{turn} = 4.5, C_{nonturn} = 1$	BTSVM (*)	78.4	55.9	58.6	0.43
	BTSVM + PSIPRED	79.5	58.5	56.9	0.44
$C_{turn} = 5.0, C_{nonturn} = 1$	BTSVM	77.4	53.3	64.3	0.43
	BTSVM + PSIPRED (*)	78.7	56.0	62.0	0.45
$C_{turn} = 5.5, C_{nonturn} = 1$	BTSVM	76.4	51.5	67.4	0.43
	BTSVM + PSIPRED	78.0	54.5	64.9	0.45
$C_{turn} = 6.0, C_{nonturn} = 1$	BTSVM	75.4	49.9	70.7	0.43
	BTSVM + PSIPRED	77.3	53.0	68.0	0.45
$C_{turn} = 7.0, C_{nonturn} = 1$	BTSVM	73.4	47.5	75.4	0.43
	BTSVM + PSIPRED	76.0	50.9	72.0	0.45

formation of  $\beta$ -turns. While amino acids Ile (I), Ala(A) and Leu (L) have a tendency to the prevention of  $\beta$ -turns; amino acid Asp (D), conversely, has a tendency to the formation of  $\beta$ -turns. Most of the other amino acids support the formation of  $\beta$ -turns when they occur at some specific positions and prevent the formation of  $\beta$ -turns when they occur at others. For example, amino acid Pro (P) strongly supports the formation of  $\beta$ -turns when it occurs at positions 4 and 7 but slightly prevents the formation of  $\beta$ -turns when it occurs at positions 2 and 5; amino acid Lys (K), when occurring at position 3, strongly supports the formation of  $\beta$ -turns, but when it occurs at position 4, strongly prevents the formation of  $\beta$ -turns; etc. These results agree closely with previously published results, which were found by statistical methods [3, 8]. Amino acid positions with the strongest supports in our results closely correspond to those with highest potentials (preferences) for  $\beta$ -turns in the previous results. Amino acid positions with the lowest supports, on the other hand, closely correspond to those with the lowest potentials (preferences). Furthermore, our results indicated that some amino acids at the both ends of the window (position 1, 2, 7 and 8), although they may not be in the  $\beta$ -turn region, also significantly support (or prevent) the  $\beta$ -turn formation of the four central residues. For example, amino acids Pro (P), Lys (K), Thr (T), and Asn (N) support the formation of  $\beta$ -turns when they occur at position 7; while amino acid Asn (D) at position 1 and Gln (Q) at position 8 prevent the formation of  $\beta$ -turns.

## 4 Discussion

### 4.1 Prediction of $\beta$ -Turns

Our method gave prediction results clearly and had better performance than other ones including BTPRED on the dataset of 426 non-homologous protein chains. The reasons for these may be the following:

1. As explained in the paper of Kaur and Raghava [12], our method, like BTPRED, has incorporated the evolutionary information of proteins by using multiple sequence alignment. The evolutionary information has been proved to significantly improve most structure prediction methods.

Table 3: The supports ( $\beta_{ai}$ ) of amino acid positions for the formation of  $\beta$ -turns (turn-windows) under the linear classification model of BTSVM.Lin. Amino acid positions with positive supports will contribute to the formation of  $\beta$ -turn, others will prevent the formation of  $\beta$ -turn. The larger the absolute value of the support, the stronger the contribution (or prevention if negative). Amino acid positions with the strongest supports (more than 0.50) are printed by boldface. Those with the lowest supports (less than -0.50) are underlined.

Amino acid	Position 1	2	3 (i)	4 (i+1)	5 (i+2)	6 (i+3)	7	8
Ala (A)	-0.346	<u>-0.539</u>	<u>-1.047</u>	0.223	<u>-0.622</u>	0.011	-0.435	-0.462
Arg (R)	-0.088	0.201	<u>-0.788</u>	0.349	0.275	0.271	0.377	-0.209
Asn (N)	-0.416	-0.164	0.122	0.400	<b>2.712</b>	0.267	<b>0.516</b>	-0.104
Asp (D)	-0.325	0.358	<b>0.589</b>	<b>0.690</b>	<b>1.542</b>	0.339	0.188	-0.367
Cys (C)	-0.131	0.138	-0.138	-0.472	0.067	0.286	0.069	0.098
Gln (Q)	-0.090	-0.227	<u>-1.140</u>	-0.083	0.106	<b>0.594</b>	0.122	-0.496
Glu (E)	-0.082	-0.286	<u>-1.242</u>	<b>0.798</b>	0.028	-0.400	0.285	-0.257
Gly (G)	-0.162	-0.044	-0.432	0.219	<b>2.207</b>	<b>1.061</b>	0.358	-0.278
His (H)	0.001	0.152	-0.412	0.250	<b>0.641</b>	0.499	0.382	0.186
Ile (I)	-0.094	-0.328	<u>-1.250</u>	-0.279	-0.489	<u>-0.702</u>	-0.202	-0.161
Leu (L)	-0.391	-0.311	<u>-0.877</u>	-0.299	-0.259	-0.242	0.002	-0.151
Lys (K)	-0.045	-0.221	<u>-1.021</u>	<b>0.944</b>	0.098	0.446	<b>0.741</b>	-0.211
Met (M)	-0.239	-0.312	<u>-0.738</u>	-0.279	-0.003	0.204	-0.009	-0.323
Phe (F)	-0.236	-0.150	<u>-0.648</u>	-0.409	0.368	-0.052	-0.048	-0.078
Pro (P)	0.077	-0.215	0.204	<b>1.982</b>	-0.254	0.263	<b>1.234</b>	0.227
Ser (S)	-0.206	-0.124	0.156	0.372	0.333	0.240	0.332	-0.282
Thr (T)	-0.214	-0.070	-0.427	-0.137	0.258	<b>0.502</b>	<b>0.724</b>	-0.052
Trp (W)	-0.263	-0.036	<u>-0.801</u>	-0.086	-0.044	-0.138	0.120	0.045
Tyr (Y)	0.177	0.089	<u>-0.511</u>	-0.164	0.230	-0.059	0.159	-0.125
Val (V)	0.034	0.044	<u>-0.911</u>	-0.326	<u>-0.542</u>	0.019	0.011	0.217

2. Like BTPRED, our method can improve the prediction accuracy by using the additional secondary structure, which is in turn predicted by a secondary structure prediction method with high accuracy, i.e. PSIPRED.
3. In our method, the prediction is performed by considering, at the same time, whether or not four consecutive residues form  $\beta$ -turn (see materials and methods). This is different from BTPRED, which performed the prediction for each residue separately. Therefore, all  $\beta$ -turns predicted by our method (containing at least four residues) are valid and clearer. There is no need to go through the filtering process to exclude unrealistic  $\beta$ -turns. In this way, our method is somewhat nearer to the nature of the problem.
4. Our method used the SVM technique, which has many advantages over the neural network technique, such as it always gives the global optimal solution with a particular kernel, is easy to control the capacity, etc. [7, 19].

## 4.2 Supports of Amino Acid Positions to the Formation of $\beta$ -Turns

We proposed the new term “support of an amino acid position to the formation of  $\beta$ -turns (turn-windows) under the SVM classification model” which emphasizes the discriminate features between  $\beta$ -turn and non- $\beta$ -turn. With some exceptions, our analysis results agree closely with those from the

previous statistical methods. That is, amino acid positions with the stronger positive supports to the formation of  $\beta$ -turns often are those with the higher amino acid positional potentials (preferences) for  $\beta$ -turns in the previous results such as those reported in [8] and Chou [3]. Conversely, amino acid positions with the stronger negative supports to the formation of  $\beta$ -turns often are those with lower amino acid positional potentials (preferences) for  $\beta$ -turns.

However, there are at least four differences between our approach and others. First, our analysis and prediction are based on the “multivariable” classification model of a SVM, which is more general than the previous models such as Site-Independent model [6], 1-4 and 2-3 Residue-Correlation model [21], and Sequence-Couple model [4]. Therefore, the supports of amino acid positions are mutually taken. This explains why the order of amino acid positions sorted by their supports is different from the order when they are sorted by their potentials (or preferences). Second, our analysis and prediction decide whether four consecutive residues are in  $\beta$ -turn or not by a window of size at least 8 (8 for BTSVM\_LIN and 12 for BTSVM with RBF kernel). Our analysis results showed that some amino acids, although may not be in  $\beta$ -turn positions, have significant supports (or preventions) to the  $\beta$ -turn formation of the residues preceding or following them. This explains why previous methods (except BTPRED) had low prediction performance, since they performed their prediction only under a window of size 4. Third, our approach has already emphasized the discriminative features of amino acid positions for the formation of  $\beta$ -turn or non- $\beta$ -turn. This is inherited from the discriminative SVM model. Finally, the analysis results of our approach are more comprehensive and therefore easier to use than those of others. Amino acid positions with positive supports will contribute to the formation of  $\beta$ -turns; otherwise they will prevent the formation of  $\beta$ -turns. The stronger the support, the stronger the affect of the amino acid position on the formation of  $\beta$ -turns.

In this research, our analysis is only taken under the linear classification model (BTSVM\_LIN) because of the complexity. As shown in Table 1, BTSVM\_LIN has the prediction performance of  $Q_{total} = 73.1$ ,  $Q_{pred} = 46.0$ ,  $Q_{obs} = 55.0$  and  $MCC = 0.32$  only. It is far from the best model for the prediction of  $\beta$ -turns, which is nonlinear model (BTSVM with RBF kernel, in our experiments). We will assess the supports of amino acid positions to the formation of  $\beta$ -turns under nonlinear classification model in future studies.

In summary, our method for predicting  $\beta$ -turns with high accuracy and our easy-to-understand analysis results are helpful to the researchers working in the fields of fold recognition and design of new molecules. Our method can be applied to prediction and analysis of the other types of tight turns.

## Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Information Science” from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. We would like to thank Harpreet Kaur for providing us some important information about BTPRED. We are also grateful to Chih Jen Lin for providing the software LIBSVM and some email discussion about it.

## References

- [1] Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z. *et al.*, Gapped Blast and PSI-Blast: a new generation of protein database search programs, *Nucl. Acids Res.*, 25:3389–3402, 1997.
- [2] Chang, C.C. and Lin, C.J., LIBSVM: a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [3] Chou, K.C., Prediction of tight turns and their types in proteins, *Analytical Biochem.*, 286:1–16, 2000.

- [4] Chou, K.C. and Blinn, J.R., Classification and prediction of  $\beta$ -turn types, *J. Protein Chem.*, 16:575–595, 1997.
- [5] Chou, P.Y. and Fasman, G.D., Conformational parameters for amino acids in helical,  $\beta$ -sheet and random coil regions calculated from proteins, *Biochemistry*, 13:211–222, 1974.
- [6] Chou, P.Y. and Fasman, G.D., Prediction of  $\beta$ -turns, *Biophys. J.*, 26:367–384, 1979.
- [7] Cristianini, N. and Shawe Taylor, J., *An Introduction to Support Vector Machines*, Cambridge, 2002.
- [8] Guruprasad, K. and Rajkumar, S.,  $\beta$ - and  $\gamma$ -turns in proteins revisited: a new set of amino acid dependent positional preferences and potential, *J. Biosci.*, 25:143–156, 2000.
- [9] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning*, 46(1/3):389–422, 2002.
- [10] Hutchinson, E.G. and Thornton, J.M., A program to identify and analyze structural motifs in proteins, *Protein Sci.*, 5:212–220, 1996.
- [11] Kaur, H. and Raghava, G.P.S., An evaluation of  $\beta$ -turn prediction methods, *Bioinformatics*, 18:1508–1514, 2002.
- [12] Kaur, H. and Raghava, G.P.S., Prediction of  $\beta$ -turns in proteins from multiple alignment using neural network, *Protein Sci.*, 12:627–634, 2003.
- [13] Kim, H. and Park, H., Protein secondary structure prediction by support vector machines and position-specific scoring matrices, *Protein Engin.*, 16(8):553–560, 2003.
- [14] Minakuchi, Y., Satou, K., and Konagaya, A., Prediction of protein-protein interaction sites using support vector machines, *Proc. International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, 22–28, 2003.
- [15] Richardson, J.S., The anatomy and taxonomy of protein structure, *Adv. Protein Chem.*, 34:167–339, 1981.
- [16] Rose, G.D., Gierasch, L.M., and Smith, J.A., Turns in peptides and proteins, *Adv. Protein Chem.*, 37:100–109, 1985.
- [17] Shepherd, A.J., Gorse, D., and Thornton, J.M., Prediction of the location and type of  $\beta$ -turns in proteins using neural networks, *Protein Sci.*, 8:1045–1055, 1999.
- [18] Takano, K., Yamagata, Y., and Yutani, K., Role of amino acid residues at turns in the conformational stability and folding of human lysozyme, *Biochemistry*, 39:8655–8665, 2000.
- [19] Vapnik, V., *Statistical Learning Theory*, Wiley N.Y., 1998.
- [20] Wilmot, C.M. and Thornton, J.M., Analysis and prediction of the different types of  $\beta$ -turns in proteins, *J. Mol. Biol.*, 203:221–232, 1988.
- [21] Zhang, C.T and Chou, K.C., Prediction of  $\beta$ -turns in proteins by 1-4 & 2-3 correlation model, *Biopolymers*, 41:673–702, 1997.