BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 10 2005, pages 2370–2374
doi:10.1093/bioinformatics/bti358

Structural bioinformatics

Improved method for predicting β-turn using support vector machine

Qidong Zhang, Sukjoon Yoon† and William J. Welsh*

Department of Pharmacology, University of Medicine and Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School and Informatics Institute of UMDNJ, 675 Hoes Lane, Piscataway, NJ 08854, USA

Received on December 31, 2004; revised and accepted on February 24, 2005
Advance Access publication March 29, 2005

ABSTRACT

Motivation: Numerous methods for predicting β-turns in proteins have been developed based on various computational schemes. Here, we introduce a new method of β-turn prediction that uses the support vector machine (SVM) algorithm together with predicted secondary structure information. Various parameters from the SVM have been adjusted to achieve optimal prediction performance.

Results: The SVM method achieved excellent performance as measured by the Matthews correlation coefficient (MCC = 0.45) using a 7-fold cross validation on a database of 426 non-homologous protein chains. To our best knowledge, this MCC value is the highest achieved so far for predicting β-turn. The overall prediction accuracy Q_{total} was 77.3%, which is the best among the existing prediction methods. Among its unique attractive features, the present SVM method avoids overtraining and compresses information and provides a predicted reliability index.

Availability: The algorithm is available via a web server on: http://serine.umdnj.edu/∼zhangq3/betaturn/

Contact: welshwj@umdnj.edu

Supplementary information: http://serine.umdnj.edu/∼zhangq3/betaturn

INTRODUCTION

Protein architecture consists of α-helices, β-sheets, tight turns, bulges and random coil structures, where the first two are repetitive motif elements and the remaining three are non-repetitive motif elements (Richardson, 1981). β-turn is a particular type of tight turn that consists of four consecutive residues which are not within an α-helix. The distance between the first and fourth (the last) C_α is <7 Å. On average, about 25% of all protein residues comprise β-turns (Kabsch and Sander, 1983).

As one of the most common types of non-repetitive motifs in proteins, β-turns bear great significance in protein structure and function. First, β-turns are four-residue reversals in proteins so that they help in the formation of higher-order structure (Takano *et al.*, 2000). Second, most β-turns are located on the surface of proteins which suggests their involvement in molecular recognition processes and interactions between receptors and substrates (Chou, 2000; Rose *et al.*, 1985). Consequently, development of an accurate prediction method of β-turns would be helpful for fold recognition studies and for predicting the overall 3D structure of proteins.

A number of β-turn prediction methods currently exist, most of which are empirical based on position preference. The present method is compared with some other β-turn prediction methods that were recently evaluated by Kaur and Raghava (2002): the Chou–Fasman method (Chou and Fasman, 1974), the 1–4 and 2–3 correlation model (Zhang and Chou, 1997), the sequence coupled model (Chou, 1997), GORBTURN (v3.0) (Gibrat *et al.*, 1987; Wilmot and Thornton, 1990) and BTPRED (Shepherd *et al.*, 1999). In the Chou–Fasman method (Chou and Fasman, 1974), a set of probabilities is assigned to each residue and the conformational parameters and positional frequencies are determined by calculating the relative frequency of each secondary structure. In the 1–4 and 2–3 correlation model (Zhang and Chou, 1997), the coupling effects between the first and fourth residues and between the second and third residues are taken into account. In the sequence coupled model developed by Chou (Chou, 1997) within the first-order Markov chain framework, the sequence correlation effect for an entire oligopeptide is considered.

GORBTURN uses the positional frequencies and equivalent parameters (Gibrat *et al.*, 1987) to remove the potential helix and strand forming residues from the β-turn prediction (Wilmot and Thornton, 1990). A neural network method, BTPRED, was developed by Shepherd *et al.* (1999) to predict the location and type of β-turns in proteins. BTPRED was found to be most accurate among these β-turn prediction methods according to side-by-side evaluation conducted by Kaur and Raghava (2002). Recently, an improved neural network method, BetaTPred2, was developed by Kaur and Raghava (2003). In this method, a great improvement in prediction performance has been achieved (Matthews correlation coefficient MCC = 0.43) by using multiple sequence alignment as input instead of the single amino acid sequence.

The present method employs a support vector machine (SVM) learning system to predict β-turns in proteins. The SVM, first proposed by Vapnik and his co-workers (Cortes and Vapnik, 1995; Vapnik, 1998), is based on statistical learning theory. Among the many attractive features of the SVM algorithm is the absence of local minima, its speed and scalability, and its ability to condense information contained in the training set. Basically, the SVM maps the input samples into feature space typical of higher dimension. Within this feature space, the SVM seeks a hyperplane [called the optimal separating hyperplane, (OSH)] that can differentiate the two classes

*To whom correspondence should be addressed.

†Present address: Department of Biological Science, Sookmyung Women's University, Seoul, Republic of Korea

with maximal margin and least error. The SVM is an extremely successful learning theory that usually outperforms other machine learning technologies such as artificial neural networks (ANNs) and nearest neighbor methods. In recent years, SVMs have performed well in diverse applications of bioinformatics including prediction of secondary structure (Hua and Sun, 2001; Ward *et al.*, 2003; Guo *et al.*, 2004), classification of protein quaternary structure (Zhang *et al.*, 2003), classification and validation of cancer tissue samples (Furey *et al.*, 2000) and prediction of T-cell epitopes (Zhao *et al.*, 2003). Cai *et al.* (2002) used the SVM approach for the prediction and classification of β-turn types with good results. Here, we introduce a novel use of the SVM approach to predict β-turns. The present SVM system together with predicted secondary structure information exhibited improved performance compared with the six other β-turn prediction methods surveyed by Kaur and Raghava (2002) in terms of various statistical figures of merit.

MATERIALS AND METHODS

Dataset

The dataset of 426 non-homologous protein chains first described by Guruprasad and Rajkumar (2000) was chosen to train and test our method. This same dataset was selected by Kaur and Raghava (2002) to evaluate the performance of six β-turn prediction methods. The structure of each protein chain in this dataset has been determined by X-ray crystallography at better than 2.0 Å resolution, and no two protein chains have >25% identity. The program PROMOTIF (Hutchinson and Thornton, 1996) was implemented to identify the observed β-turns in these crystal structures.

Design

The SVMlight program was used to train the SVM classifier (Joachims, 1999). First, using the classical local coding scheme of the protein sequences with a sliding window, the amino acid type of each residue is encoded into a length-20 vector by the unary encoding scheme. Following this scheme, alanine is represented as (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0). The 'null' residue, represented by an all-zero length-20 vector, was used to fill in the empty position. Therefore, a protein fragment of window size m is represented by a $20 \times m$ matrix of zeros and ones. Second, with multiple alignments, we use the position-specific scoring matrix generated by PSI-BLAST as input to our SVM classifier. These profiles were scaled to 0–1 range using the standard logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)}, \tag{1}$$

where x is the raw profile matrix value which represents the likelihood of that particular residue substitution at that position. The structure of the multiple alignment SVM system is illustrated in Figure 1. The predicted secondary structure from PSIPRED (Jones, 1999) is encoded as follows: helix → (1,0,0), strand → (0,1,0), coil → (0,0,1). The window size was set to 7-residues in accordance with Shepherd *et al.* (1999) who found that BTPRED achieved optimal β-turn prediction with a window size of 7 or 9. Furthermore, a 7-residue sequence context is the minimum size sufficient to account for the coupling effect between the first and fourth residues within β-turn sequences.

Training and testing

We employed 7-fold cross validation to evaluate the performance of the present method. The 426 protein chains were divided into 7 subsets of equal size (i.e. 6 subsets contained 61 chains; 1 subset contained 60 protein chains). At each step of the validation process, six subsets were used for training while the remaining one subset was used for testing. This procedure was repeated seven times, once for each subset. Several parameters were adjusted for optimal performance. Our SVM employed the radial basis function kernel [Equation (2)] with a soft margin, thus the first parameters to be determined are γ and the regularization parameter C. The percentage of β-turn residues

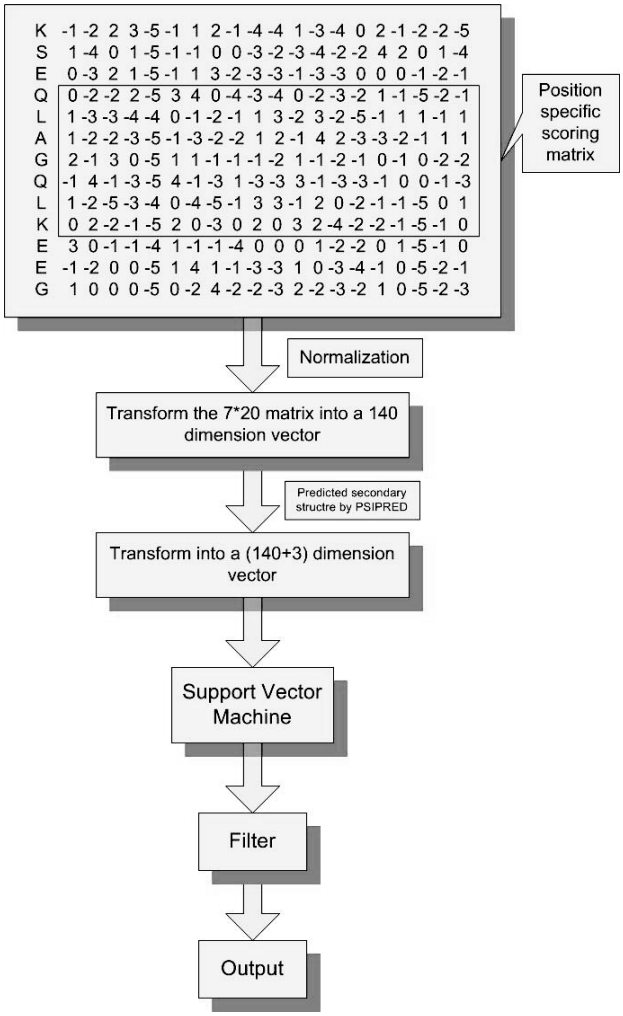

Fig. 1. The architecture of the present SVM system using multiple alignment. The protein sequence is represented by the PSI-BLAST profile and transformed into a number of 20×7 dimension vectors using the sliding-window method. After normalization, these vectors are transformed into a number of 143D vectors with predicted secondary structures and serve as inputs to the SVM.

in our dataset is roughly the same as that found (25%) in naturally occurring proteins; thus the cost factor j is used to minimize false negatives. In the present case, we set $\gamma = 0.0186$, $C = 16$ and $j = 2$. Additional information about parameter selection can be found in the Supplementary material.

$$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left|\vec{x}_i - \vec{x}_j\right|^2). \tag{2}$$

Reliability index

It is important to know the prediction reliability of machine learning techniques applied in computational biology. Here, the reliability index (RI) was used to determine the effectiveness of β-turn prediction. In addition, key regions with high prediction accuracy can be easily identified by means of RI. An intuitive RI can be derived using the output of the SVM classifier (Hua and Sun, 2001) which is a real number usually between -2 and $+2$. A sample with large positive output value is indicative of a large positive distance to the OSH and, accordingly, will have high probability of being β-turn. The RI can be defined as:

$$\mathrm{RI} = \mathrm{int}\frac{\mathrm{abs}(D)}{0.2}, \tag{3}$$

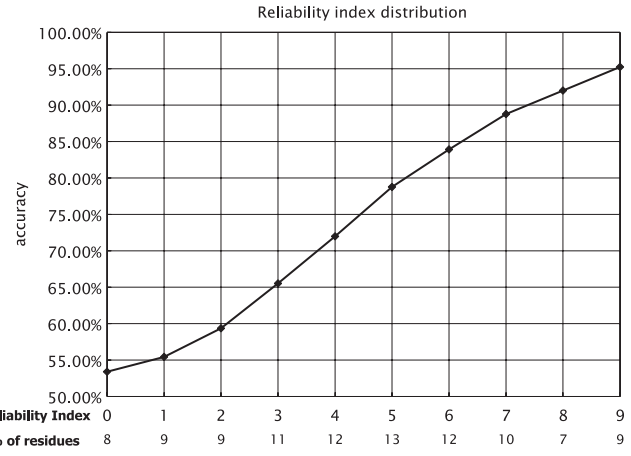

Fig. 2. Expected prediction accuracy for residues with different reliability indices. The accuracy and the fraction of residues with particular RI are given. The expected accuracy of residues with higher RI is much better than those with lower RI.

where abs(D) is the absolute value of distance D between the sample and the OSH. RI is an integer between [0, 9] where the maximal RI = 9 indicates a very reliable prediction. Figure 2 shows that the prediction is more reliable as RI increases, confirming that the RI as defined here reflects the prediction reliability.

Filtering

The prediction for each residue is made without reference to the prediction status of neighboring residues; thus the predictions are not correlated. To ensure that β-turns are at least four residues long, we added a simple filtering step known as the 'state-flipping' rule first described by Shepherd *et al.* (1999).

Performance measures

A variety of statistical measures are available to evaluate the performance of predictive methods in biology. Four measures widely used in β-turn prediction methods are based on the following scalar quantities:

(1) p, the number of correctly classified β-turn residues

(2) n, the number of correctly classified non-β-turn residues

(3) o, the number of incorrectly classified β-turn residues

(4) u, the number of incorrectly classified non-β-turn residues and

(5) t, the total number of residues.

The first measure is Q_{total} which calculates the percentage of residues that are correctly classified:

$$Q_{\text{total}} = \frac{p+n}{t} \times 100. \tag{4}$$

It is the most common measure of a method's overall performance; however, Q_{total} can be misleading as β-turn residues occur much less frequently than non-β-turn residues in proteins (~25 versus ~75%). Therefore, one could easily achieve $Q_{\text{total}} = 75\%$ merely by predicting all residues to be non-β-turn. For this reason, we calculated $Q_{\text{predicted}}$, the percentage of correctly predicted β-turns:

$$Q_{\text{predicted}} = \frac{p}{p+o} \times 100 \tag{5}$$

and Q_{observed}, the percentage of observed β-turns that are correctly predicted:

$$Q_{\text{observed}} = \frac{p}{p+u} \times 100 \tag{6}$$

$Q_{\text{predicted}}$ and Q_{observed} represent measures of the method's sensitivity and selectivity, respectively.

We also computed the MCC as a measure of both sensitivity and selectivity:

$$\text{MCC} = \frac{(p \times n) - (o \times u)}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}. \tag{7}$$

Another important consideration is whether the present method performs better than random prediction. We first calculated R, the anticipated number of residues that are correctly classified by random prediction (Shepherd *et al.*, 1999):

$$R = \frac{(p+o)(p+u) + (n+u)(n+o)}{t}. \tag{8}$$

We then calculated S, the normalized percentage of correctly predicted samples better than random:

$$S = \frac{(p+n) - R}{t - R} \times 100. \tag{9}$$

Accordingly, $S = 100\%$ for perfect prediction and $S = 0\%$ for worse than random prediction.

RESULTS AND DISCUSSION

Results from the present SVM method using single amino acid sequence as input are compared in Table 1 with BTPRED (Shepherd *et al.*, 1999) and other popular β-turn prediction methods. BTPRED, based on neural networks, is generally considered among the most reliable and accurate β-turn prediction methods. It is seen that the MCC is appreciably higher for the present method (0.41) than for BTPRED (0.35). This is noteworthy in that the MCC is a robust and reliable performance measure that accounts for both overpredictions and underpredictions. Prediction coverage Q_{observed} by the present method (67.9%) exceeds BTPRED (48%) by almost 20%. Moreover, the value of S [Equation (9)] for our method is 40% which denotes much better than random prediction.

A further improvement has been achieved by using PSI-BLAST generated scoring matrices as input (Table 2). Use of multiple alignment information reaches MCC of 0.45 and overall accuracy of 77.3%, which are best among current β-turn prediction methods (Table 3). The final SVM classifier yields $Q_{\text{predicted}}$ of 53.1% and S of 44%, which is slightly better than that of the single sequence. In conclusion, the prediction performance of our method has been further improved by using the multiple alignment information in the form of the PSI-BLAST position-specific matrices as input.

Some of the protein chains in our dataset may be used to train PSIPRED. In order to cross-validate the results, we have excluded those proteins from the non-redundant database of PSIPRED. As shown in Table 3, the difference in prediction performance is negligible.

Three factors may account for the exceptional performance of the present method. First, a new statistical learning algorithm, SVM, is employed. Among its many unique features, SVM can handle large datasets and exhibits remarkable resistance to overfitting. SVMs condense information in the training set by using a very small number of samples with support vectors (SVs) to provide sparse representation. It is believed that these SVs contain all the information needed for classification. In most cases the number of SVs is much smaller than the total number of training samples, such that the SVM can efficiently classify new samples by safely ignoring the training samples judged as unnecessary. In our method, the ratio of SVs to training samples is 55.6%, which means nearly 44.4% of the training samples could be safely removed. That SVMs can effectively remove the uninformative patterns in the dataset and focus on the informative patterns is a major asset. Second, predicted secondary structure

Table 1. Performance comparison between the present method (single sequence) and other popular methods

Methods	Q_{total}	$Q_{predicted}$	$Q_{observed}$	MCC	S
Present method (single sequence)	74.8	49.1	67.9	0.41	40%
BTPRED[a]	74.9	55.3	48.0	0.35	35%
Chou–Fasman[b]	65.2	37.6	63.5	0.26	—
1–4 and 2–3 correlation model[b]	59.1	32.4	61.9	0.17	—
Sequence coupled model[b]	53.3	32.4	72.8	0.17	—
GORBTURN[b]	70.5	39.3	37.3	0.19	—

The results of the present method using single amino acid sequence as input were obtained by a 7-fold cross validation.
— Result cannot be determined from the paper.
[a]Results obtained on another non-homologous dataset which contains 300 protein chains (Shepherd *et al.*, 1999).
[b]Results obtained on the same 426 non-homologous dataset (Kaur and Raghava, 2002).

Table 2. Prediction results using single sequence and multiple alignment

	Single sequence	Multiple alignment
Q_{total}	74.8	77.3
$Q_{predicted}$	49.1	53.1
$Q_{observed}$	67.9	67.0
MCC	0.41	0.45
S	40%	44%

Table 3. Performance comparison between the present method (multiple alignment) and the current best method, BetaTPred2

Method	Q_{total}	$Q_{predicted}$	$Q_{observed}$	MCC
Present method (multiple alignment)	77.3 (77.3)	53.1 (53.1)	67.0 (67.3)	0.45 (0.45)
BetaTPred2	75.5	49.8	72.3	0.43

Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

information by PSIPRED is used. It is widely believed that β-turn prediction accuracy can be greatly improved by inclusion of secondary structure information (Kaur and Raghava, 2002). PSIPRED, based on neural network evaluation of PSI-BLAST generated profiles (Jones, 1999), is one of the most accurate secondary structure prediction methods. Third, multiple alignment information in the form of PSI-BLAST profiles has been used as input to the SVM classifier. These profiles are generated by searching remote homologs against a huge nonredundant database and contain evolutionary information. With multiple alignment, the MCC value is raised from 0.41 to 0.45, which is the best value for β-turn prediction achieved so far.

Our method of β-turn prediction can be further improved in future work. By analyzing sequence–structure relationships in terms of tertiary contact (TC), Yoon and Welsh (2004) have successfully detected nonnative sequence propensity for amyloid fibril formation. TCs, formed during the protein folding process, are interactions between non-adjacent residues which are far apart along the first-order amino acid sequence. TC counting is an easy and fast way to quantify the influence of tertiary environment and has shown its ability to measure tertiary interactions and solvent accessibility. It is our hope that incorporation of tertiary contacts in our SVM method will yield ever higher prediction accuracy. As a passive learning machine method, SVM might be improved by combination with active learning methods such as boosting. An active learning method could directly select a subset of samples for training and testing, thereby improving the accuracy of any given passive learning algorithm. Such practical strategies that fuse different techniques to improve β-turn prediction performance are currently under development in our laboratory.

ACKNOWLEDGEMENT

The authors acknowledge the support for this research provided by a High Technology Workforce Excellence grant sponsored by the Commission on Higher Education of the State of New Jersey.

REFERENCES

Cai,Y.D. *et al.* (2002) Support vector machines for the classification and prediction of beta-turn types. *J. Pept. Sci.*, **8**, 297–301.
Chou,K.C. (1997) Prediction of beta-turns. *J. Pept. Res.*, **49**, 120–144.
Chou,K.C. (2000) Prediction of tight turns and their types in proteins. *Anal. Biochem.*, **286**, 1–16.
Chou,P.Y. and Fasman,G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.
Cortes,C. and Vapnik,V. (1995) Support vector networks. *Machine Learning*, **20**, 273–293.
Furey,T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
Gibrat,J.-F. *et al.* (1987) Further development of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, **198**, 425–433.
Guo,J. *et al.* (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles. *Proteins*, **54**, 738–743.
Guruprasad,K. and Rajkumar,S. (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.
Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.
Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT-Press, Cambridge, MA, USA.
Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

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

Kaur,H. and Raghava,G.P. (2002) An evaluation of beta-turn prediction methods. *Bioinformatics*, **18**, 1508–1514.

Kaur,H. and Raghava,G.P. (2003) Prediction of β-turns in proteins from multiple alignment using neural network. *Protein Sci.*, **12**, 627–634.

Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein. Chem.*, **34**, 167–339.

Rose,G.D. *et al*. (1985) Turns in peptides and proteins. *Adv. Protein Chem.*, **37**, 1–109.

Shepherd,A.J. *et al*. (1999) Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Sci.*, **8**, 1045–1055.

Takano,K. *et al*. (2000) Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry*, **39**, 8655–8665.

Vapnik,V. (1998) *Statistical Learning Theory*. In Schölkopf,B., Burges,C. and Smola,A. (eds), John Wiley and Sons, Inc., NY.

Ward,J.J. *et al*. (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650–1655.

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

Yoon,S. and Welsh,W.J. (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci.*, **13**, 2149–2160.

Zhang,C.T. and Chou,K.C. (1997) Prediction of beta-turns in proteins by 1–4 & 2–3 Correlation Model. *Biopolymers*, **41**, 673–702.

Zhang,S.W. *et al*. (2003) Classification of protein quaternary structure with support vector machine. *Bioinformatics*, **19**, 2390–2396.

Zhao,Y. *et al*. (2003) Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, **19**, 1978–1984.