ResearchGate

See discussions, stats, and author profiles for this publication at: http://www.researchgate.net/publication/5419532

Using support vector machine to predict β- and γ-turns in proteins

ARTICLE *in* JOURNAL OF COMPUTATIONAL CHEMISTRY · MAY 2008

Impact Factor: 3.6 · DOI: 10.1002/jcc.20929 · Source: PubMed

CITATIONS
26

DOWNLOADS
8

VIEWS
55

2 AUTHORS, INCLUDING:

Qianzhong Li
Inner Mongolia University
57 PUBLICATIONS **1,030** CITATIONS

SEE PROFILE

Available from: Qianzhong Li
Retrieved on: 08 August 2015

Using Support Vector Machine to Predict β- and γ-Turns in Proteins

XIUZHEN HU,[1,2] QIANZHONG LI[1]
[1]*Laboratory of Theoretical Biophysics, Department of Physics, College of Sciences and Technology, Inner Mongolia University, Hohhot, People's Republic of China*
[2]*Department of Physics, College of Sciences, Inner Mongolia University of Technology, Hohhot, People's Republic of China*

Received 17 September 2007; Revised 22 December 2007; Accepted 27 December 2007
DOI 10.1002/jcc.20929
Published online 23 April 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: By using the composite vector with increment of diversity, position conservation scoring function, and predictive secondary structures to express the information of sequence, a support vector machine (SVM) algorithm for predicting β- and γ-turns in the proteins is proposed. The 426 and 320 nonhomologous protein chains described by Guruprasad and Rajkumar (Guruprasad and Rajkumar J. Biosci 2000, 25,143) are used for training and testing the predictive model of the β- and γ-turns, respectively. The overall prediction accuracy and the Matthews correlation coefficient in 7-fold cross-validation are 79.8% and 0.47, respectively, for the β-turns. The overall prediction accuracy in 5-fold cross-validation is 61.0% for the γ-turns. These results are significantly higher than the other algorithms in the prediction of β- and γ-turns using the same datasets. In addition, the 547 and 823 nonhomologous protein chains described by Fuchs and Alix (Fuchs and Alix Proteins: Struct Funct Bioinform 2005, 59, 828) are used for training and testing the predictive model of the β- and γ-turns, and better results are obtained. This algorithm may be helpful to improve the performance of protein turns' prediction. To ensure the ability of the SVM method to correctly classify β-turn and non-β-turn (γ-turn and non-γ-turn), the receiver operating characteristic threshold independent measure curves are provided.

© 2008 Wiley Periodicals, Inc. J Comput Chem 29: 1867–1875, 2008

Key words: increment of diversity; β-turn; γ-turn; position conservation scoring function; support vector machine

Introduction

Protein secondary structure prediction is a key step for predicting tertiary structure of proteins. In the past three decades, a large number of methods have been developed for predicting the regular secondary structures (α-helix, β-strand) and coil in proteins.[1–6] However, they could not provide a directional change for the polypeptide chain. Therefore, the prediction of tight turns in proteins is as important as helix and strand prediction. The tight turns play an important role in protein, such as folding stability,[7,8] recognition,[9,10] and structure assembly.[11]

Tight turns can be divided into δ-, γ-, β-, α-, and π-turns according to the number of residues involved.[12] The β-turns are the most common turns in proteins. A β-turn comprises four consecutive residues, which does not form α-helix, and the distance between $C_\alpha(i)$ and $C_\alpha(i + 3)$ is less than 7 Å. According to backbone dihedral angles in the inner residues, $i + 1$ and $i + 2$ will define different types of β-turns.

Because of β-turn's ability to reverse the direction of a protein chain to 180°, it is responsible for the compact globular shape of proteins. The β-turn is an important component of β-hairpin structure and plays a vital role in protein folds. Enhancing β-turn prediction can have a direct effect on molecular recognition studies and the identification of important structural motifs, such as β-hairpins. It also contributes indirectly to the overall prediction of protein tertiary structures.

Some methods have been developed for prediction of β-turns based on the statistical model and machine learning technique.[13–22] The typical β-turns prediction was made by Kaur and Raghava.[13] They used the same dataset with 426 protein chains constructed by Guruprasad and Rajkumar[23] and the same measures, compared with some other β-turn prediction methods, such as Chou-Fasman,[24] the 1–4 and 2–3 correlation model,[18] the sequence coupled model,[17] GORBTURN (v3.0),[25,26] and

Correspondence to: Q.-Z. Li; e-mail: qzli@imu.edu.cn

Contract/grant sponsor: National Natural Science Foundation of China-contract/grant number: 30560039

Contract/grant sponsor: Natural Science Foundation of the Inner Mongolia of China; contract/grant numbers: 200508010509, 200607010101.

Contract/grant sponsor: Project for Excellent Subject-Directors of Inner Mongolia Autonomous Region

© 2008 Wiley Periodicals, Inc.

BTPRED.[16] In addition, an improved neural network method, BetaTPred2, was developed by Kaur and Raghava.[14] In this method, a great improvement in prediction performance was that the Matthews correlation coefficient (Mcc) = 0.43 had been achieved by using the multiple sequence alignment as input instead of the single amino acid sequence. Farther, Cai et al.[19] and Lin et al.[20] used the support vector machine (SVM) and the Markov Chains theory to predict β-turns, and obtained significant results. Fuchs and Alix[22] predicted β-turns in the dataset of 426 proteins using propensities and multiple alignments. The obtained Mcc and overall prediction accuracy were 0.42 and 74.8%, respectively. The better prediction accuracy based on SVM was obtained using input parameters of the predicted secondary structure and multiple alignment information among these methods.[21,27] The overall prediction accuracy and Mcc in the 7-fold cross-validation were 77.3% and 0.45,[21] and 79.8% and 0.45,[27] respectively, for 426 nonhomologous protein chains described by Guruprasad and Rajkumar.[23]

Recently, the different kinds of turns were studied by Street et al.[28] Their results provided a molecular explanation for the observation that reverse turns between elements of regular secondary structure can be classified into a small number of discrete conformations. And Bornot and de Brevern[29] analyzed the distributions of β-turns according to different secondary structure assignment methods.

The γ-turn is the second most characterized and commonly found turn, after the β-turn. A γ-turn is defined as a three-residue turn with a hydrogen bond between the carbonyl oxygen of residue i and the hydrogen of the amide group of residue $i + 2$. There are two types of γ-turns: inverse and classic.[30]

γ-Turns also play an important role in protein folding and recognition. Experimentation indicated that some folds are achieved by γ-turn reverse,[31] and some acceptor combining sites are also in the γ-turn.[32] Although γ-turn content is fewer, it contained important information on the molecule recognition. Therefore, based on the amino acid sequence, predicting γ-turns is significant.

Compared with β-turns, however, γ-turns are seldom investigated. This is because of the lower occurrence of γ-turns in proteins. On the basis of multiple alignment and predicted secondary structure information, the γ-turns in proteins were predicted by Kaur and Raghava using neural network method.[33] The Mcc was 0.17 in the 5-fold cross-validation and the corresponding area under the receiver operating characteristic (ROC) curves was 0.73 for 320 nonhomologous protein chains described by Guruprasad and Rajkumar.[23] Guruprasad et al.[23,34] used the Markov Chains theory to predict γ-turns, and obtained significant results. Pham et al.[27] used SVM method at the residue level and turn level to predict γ-turns. The Mcc was 0.13 in the 5-fold cross-validation, and the overall prediction accuracy was 79.9% for 320 nonhomologous protein chains.

In this study, based on the datasets described by Guruprasad and Rajkumar[23] and by Fuchs and Alix,[22] by using the composite vector at the residue level, including the ID'PCSF values and predictive secondary structure information as inputting parameters of the SVM, the LIBSVM program packages are applied to predict β- and γ-turns, respectively. More accurately predicted results are obtained.

Materials and Methods

Materials

The two datasets described in the work of Guruprasad and Rajkumar[23] are used. A total of 426 and 320 nonhomologous protein chains with resolution <2.0 Å and sequence identity <25% are used in our method. Three hundred and fifteen out of 320 protein chains are contained in the above 426 protein chains. Four hundred and twenty six and 320 nonhomologous protein chains are respectively used for the prediction of β- and γ-turns.

In addition, the two datasets described in the work of Fuchs and Alix[22] are used. Among a total of 547 and 823 nonhomologous protein chains with a resolution of <2.0 Å and a sequence identity of <25%, each chain contains one minimum β-turn, 543 and 819 proteins that contained respectively 7912 and 11257 β-turns, and are used for β-turns prediction. The 346 and 536 nonhomologous protein chains in two datasets, each chain containing one minimum γ-turn, are used for γ-turns prediction. They contain 873 and 1303 γ-turns, respectively.

Two hundred and ten out of 543 protein chains are contained in the above 819 protein chains. One hundred and ninety one out of 543 protein chains are contained in the above 426 protein chains. Ninety out of 819 protein chains are contained in the above 426 protein chains.

The secondary structure was assigned to each amino acid of two protein datasets by using DSSP.[35] The program PROMOTIF[36] was implemented to identify the observed β- and γ-turn motifs.

Methods

Position Conservation Scoring Function

The position conservation scoring function (PCSF) method had been widely used in the prediction of transcription factor binding sites in genomes.[37–41] To consider the effect of position in β- or γ-turn sequence segments, PCSF will be constructed by calculating the position probability matrix (PPM) and the conservation index of position.

PPM and Conservation Index of Position

The PPM includes $20 \times L$ elements (L is the length of the sequence segments, 20 denotes the 20 native amino acids). Each element in the matrix represents probability at a corresponding position, which is defined as:

$$P_{i,x} = \frac{n_{i,x} + s_x}{N_i + \sum_{x=1}^{20} s_x} \quad (1)$$

where, $n_{i,x}$ and s_x, respectively, denote the real counts and pseudo counts for amino acid x at the i-th position of the sequence segments. s_x is calculated by[37,38]:

$$s_x = \frac{\sqrt{N_i}}{20} \quad (2)$$

where N_i is the total number of the sequences.

To reflect the action of position information in the sequence segment, the conservation index at the i-th position may be defined by the following expressions[39,40]:

$$c_i = \frac{100}{\log 20}\left(\sum_{x=1}^{20} p_{i,x}\log p_{i,x} + \log 20\right) \quad (3)$$

$p_{i,x}$ is defined by equation (1), c_i value equals 100 for full conservation at the i-th position, c_i equals 0 for random amino acids at the i-th position. The conservation index of position in the amino acid sequence reflects the difference of amino acid compositions in same position between different datasets (i.e. between β-turn and non-β-turn datasets, or between γ-turn and non-γ-turn datasets).

Scoring Function

For an arbitrary sequence segment S with N amino acids (i.e. $S = (x_1, x_2,.....,x_N)$, where x_i is the amino acid at position i in segment S), the score of segment S can be defined as[39,41]:

$$F(S) = F(x_1, x_2, ..., x_N) = \frac{\sum_{i=1}^{N} c_i\left(p_{i,x} - p_{i,x}^{\min}\right)}{\sum_{i=1}^{N} c_i\left(p_{i,x}^{\max} - p_{i,x}^{\min}\right)} \quad (4)$$

It is easily proven that $0 \le F(S) \le 1$. Here, $p_{i,x}^{\min} = \min(p_{i,x})$ and $p_{i,x}^{\max} = \max(p_{i,x})$ are the minimal and maximal values of amino acid probabilities at position i, respectively.

For an arbitrarily sequence segment, the class of this segment may be predicted by the maximum among $F(S)^{\beta}$ and $F(S)^{\text{non-}\beta}$ ($F(S)^{\gamma}$ and $F(S)^{\text{non-}\gamma}$), and can be formulated as follows:

For prediction β-turns: $F(S)^{\xi} = Max\ \{F(S)^{\beta}, F(S)^{\text{non-}\beta}\}$
$\zeta \in \beta$-turn or non-β-turn.
For prediction γ-turns: $F(S)^{\xi} = Max\ \{F(S)^{\gamma}, F(S)^{\text{non-}\gamma}\}$
$\zeta \in \gamma$-turn or non-γ-turn.

The operator Max means taking the maximum value among those in the parentheses, and then the ξ will give the segment class to which the predicted segment should belong (i.e. center amino acid of segment is should belong).

Increment of Diversity

The increment of diversity (ID) algorithm, a whole uncertain measure and total information of a system on state space, is essentially a measure of the composition similarity level for two systems.[42] The ID algorithm has been applied in the recognition of protein structural class,[43] the exon–intron splice site prediction,[44] and the prediction of subcellular location of proteins.[45]

In the state space of t dimension, the diversity measure for diversity source S: $\{m_1, m_2,\ldots, m_t\}$ is defined as[42–45]:

$$D(S) = M\log M - \sum_i m_i \log m_i \quad (5)$$

In the same state space, ID between the source of diversity $X(n_1, n_2, ...,n_t)$ and $S(m_1, m_2, ...,m_t)$ is defined as:

$$\mathrm{ID}(X,S) = D(X+S) - D(X) - D(S) \quad (6)$$

It is easily proven that the ID can be written as:

$$\mathrm{ID}(X,S) = D(M,N) - \sum_i D(m_i,n_i) \quad (7)$$

$$D(M,N) = (M+N)\log(M+N) - M\log M - N\log N \quad (8)$$

$$D(m_i,n_i) = (m_i+n_i)\log(m_i+n_i) - m_i\log m_i - n_i\log n_i \quad (9)$$

here $N = \sum_i n_i, M = \sum_i m_i$. If m_i or n_i equals zero, then $D(m_i, n_i) = 0$.

An arbitrarily sequence segment may be predicted by the minimum among ID^{β} and $\mathrm{ID}^{\text{non-}\beta}$ (ID^{γ} and $\mathrm{ID}^{\text{non-}\gamma}$), and can be formulated as follows:

For prediction β-turns: $\mathrm{ID}^{\xi} = \mathrm{Min}\{\mathrm{ID}^{\beta}, \mathrm{ID}^{\text{non-}\beta}\}$
$\zeta \in \beta$-turn or non-β-turn.
For prediction γ-turns: $\mathrm{ID}^{\xi} = \mathrm{Min}\{\mathrm{ID}^{\gamma}, \mathrm{ID}^{\text{non-}\gamma}\}$
$\zeta \in \gamma$-turn or non-γ-turn.

The operator Min means taking the minimum value among those in the parentheses, and then the ξ will give the segment class to which the predicted segment should belong (i.e. center amino acid of segment is should belong). This method is also applied to predict γ-turns in the proteins.

Support Vector Machine

The SVM is an extremely successful learning system. SVM has been widely used in pattern recognition[46,47] and prediction of the secondary structure in proteins.[5,48] In this article, LIBSVM software packages are used.[49]

Parameter Selection

The probabilities of 20 amino acids at each position are important parameters for predicting β- and γ-turns and have been introduced by the previous investigators.[13,14,18,27] They are selected as parameters (A_p) of PCSF, using the training dataset to construct PPMs (contained $20 \times L$ elements). For every sequence segment, the 2 score can be obtained for the β-turn and non-β-turn (γ-turn and non-γ-turn).

In the ID algorithm, D_0, C_0, and H_2 are, respectively, selected as the parameters of diversity source of the β-turns and non-β-turns. The D_0 denotes the frequencies of 400 dipeptide compositions from 20 amino acids; C_0 denotes the frequencies of six hydropathy characteristics at each position; and H_2 denotes dipeptide frequencies of the residue i and reduces $i + 3$ for the six hydropathy characteristics. They constructed 20×20, 6×7, and 6×6 dimensions of state space, respectively. For example, D_0 is selected as the parameters of diversity source in equation (5), based on the equation (7) definition, for all sequence segments, the 2 ID can be obtained for the β-turns and non-β-turns.

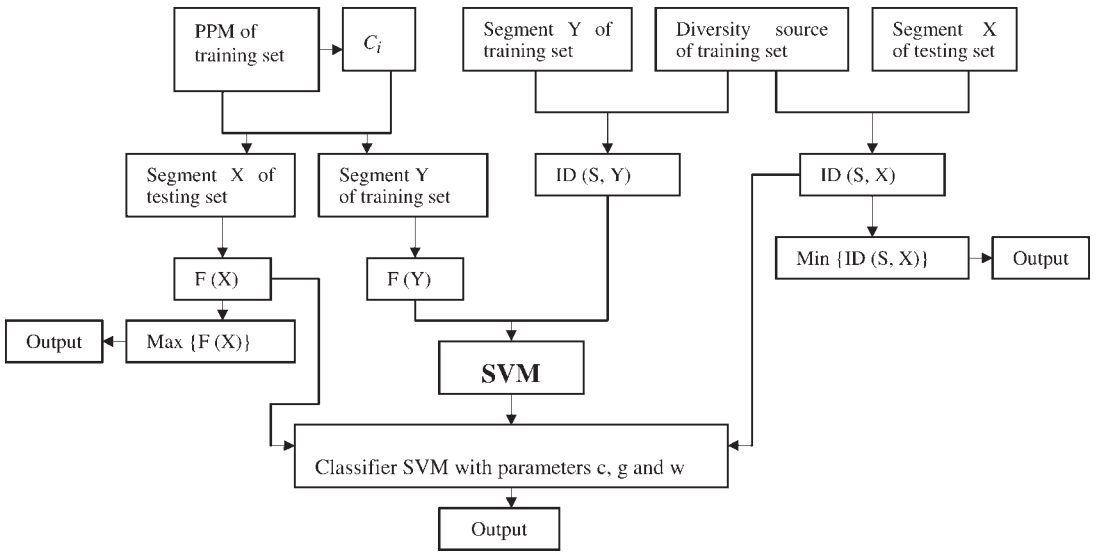

Figure 1. The architecture of the turn prediction system. Note: The architecture contains the PCSF prediction system, the ID algorithmic prediction system, and the SVM prediction system. The C_i is the conservation index at the i-th position; the ID is the increment of diversity; the F is the scoring function.

A_0 and H_0 are, respectively, selected as the parameters of diversity source of the γ-turns and non-γ-turns. A_0 denotes the frequencies of 20 native amino acids at each position; H_0 denotes the frequencies of contiguous residue dipeptide compositions from six hydropathy characteristics. They constructed the 20×5 and 6×6 dimensions of state space, respectively.

When using SVM algorithm to predict β-turn and non-β-turn, for every sequence segment, the 2 scores can be obtained by the PCSF algorithm, and 6 ID (for D_0, C_0, and H_2 parameters) values can be obtained by ID algorithm. The 8 parameters are selected as input parameters of the SVM.

For γ-turn and non-γ-turn prediction, for every sequence segment, the 2 scores can be obtained by the PCSF algorithm, and 4 ID (for A_0 and H_0 parameters) values can be obtained by ID algorithm. The 6 parameters are selected as input parameters of the SVM.

The architecture of the turn prediction system is shown in Figure 1.

Filtering and k-Fold Cross-Validation

In the case of predicting β-turns (γ-turns) at the residue level, the sliding windows contain 7 (5) amino acids, and the center amino acid of the segment is a predicted amino acid. To ensure prediction veracity, the structure characteristic of the β-turn (γ-turn) that contained 4 (or 3) consecutive residues is considered.[16]

To compare our method with other methods, the same 7-fold cross validation technique used by Kaur and Raghava[13,14] and Zhang et al.[21] for predicting β-turns is used. The 426 proteins are randomly divided into 7 subsets (6 subsets contained 61 chains; 1 subset contained 60 protein chains). Also, the 543 proteins are randomly divided into 7 subsets (3 subsets contained 77 chains; 4 subsets contained 78 protein chains). The 819 proteins are randomly divided into 7 subsets (7 subsets contained 117 chains). Each subset is an unbalanced set that retains the naturally occurring proportion of β-turns and non-β-turns (1:3). The methods were trained on 6 subsets, and the performance is measured on the remaining seventh subset. This process is repeated 7 times so that each subset is tested.

Similarly, for the γ-turns prediction, we have used the 5-fold cross validation technique, which is used by Kaur and Raghava.[33] The 320, 346, and 536 proteins are randomly divided into 5 subsets, respectively: 5 subsets containing 64 chains, 4 subsets containing 69 chains, 1 subset containing 70 protein chains, 4 subsets containing 107 chains, and 1 subset containing 108 protein chains. Each subset is an unbalanced set that retains the naturally occurring proportion of γ-turns and non-γ-turns (1:28∼30), (1:30∼33), and (1:26∼30).

Performance Measures and Threshold Independent Measures

The performance of β- and γ-turns prediction is estimated by four parameters: the overall prediction accuracies (Q_{total}); percentages of correctly predicted β- or γ-turns (Q_{pred}); percentages of observed β- or γ-turns that are correctly predicted (Q_{obs}); and Mcc, calculated by:

$$Q_{\text{total}} = \frac{p+n}{p+n+u+o}$$

$$Q_{\text{pred}} = \frac{p}{p+o}$$

$$Q_{\text{obs}} = \frac{p}{p+u}$$

$$\text{Mcc} = \frac{(p \times n) - (u \times o)}{\sqrt{(p + o) \times (p + u) \times (n + o) \times (n + u)}}$$

Table 1. The Classifications of Amino Acids.

Classification	Amino Acids
Strongly hydrophilic or polar	**R, D, E, N, Q, K, H**
Strongly hydrophobic	**L, I, V, A, M, F**
Weakly hydrophilic or weakly hydrophobic	**S, T, Y, W**
Proline	**P**
Glycine	**G**
Cysteine	**C**

where p is the number of correctly classified β-turn (or γ-turn) residues, n is the number of correctly classified non-β-turn (non-γ-turn) residues, o is the number of non-β-turn (non-γ-turn) residues incorrectly classified as β-turn (γ-turn) residues, and u is the number of β-turn (γ-turn) residues incorrectly classified as non-β-turn (non-γ-turn) residues.

For a prediction method, it is important to know the prediction reliability. The area under the ROC curve used by Kaur and Raghava[13,14,33] is taken as a reliable index because it provides the effectiveness of β-turns (γ-turns) prediction. The measure of overall accuracy is not dependent on a particular threshold.[50] In the ROC plot, all sensitivity values (true positive fraction) for all available thresholds are displayed on the y-axis, and all (1-specificity) values (false-positive fraction) for all available thresholds are shown on the x-axis. Sensitivity (S_n) and specificity (S_p) are defined as:

$$S_n = p/(p+u) \qquad S_p = n/(n+o)$$

The Hydropathy Distribution along Protein Sequence

The hydropathy distribution along the protein sequence has been recognized as a feature useful for the characterization of protein structure in the form of hydropathy profiles.[51] To obtain the hydropathy characteristics, the amino acids may be divided into groups using their individual hydropathy according to the ranges of the hydropathy scale. Because Proline, Glycine, and Cysteine have unique backbone properties, they are classified into 3 groups. Therefore, a protein sequence with 20 amino acids can be represented by a sequence with 6 characters.[45] The classification of amino acids are shown in Table 1.

Results and Discussion

Prediction of β-Turns in the 426 Chains Dataset

The Predictive Results by Using PCSF Algorithm

Here the length of windows is 7 amino acids, so the PPMs contained 20 × 7 elements. The predicting results of β-turns and non-β-turns are shown in Table 2. The Mcc value is 0.26. To compare our method with other methods, other method's results for using the same dataset are also shown in Table 2.

The Predictive Results by Using ID Algorithm

Here D_0, C_0, and H_2 are selected as the parameters of the diversity source, respectively. For each kind of parameter, the performance of ID method is also shown in Table 2. Similar results are obtained by the three kinds of parameters, respectively. The Mcc achieved 0.29, which is slightly better than 0.26 of the

Table 2. The Predictive Results of Different Methods for the β-Turns in the 426 Proteins Using the 7-Fold Cross-Validation.

Method (parameter)	M	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	Mcc
PCSF (A_p)	20 × 7	75.7	55.2	32.4	0.26
ID (D_0)	400	67.8	43.2	69.7	0.29
ID (C_0)	6 × 7	66.4	43.2	67.0	0.28
ID (H_2)	36	67.4	43.3	68.2	0.29
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2))	4 × 2	77.3	54.3	67.9	0.45
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2) + PSI)	**4 × 2 + 3**	**79.8 (79.3)**	**55.6 (55.4)**	**68.9 (68.9)**	**0.47 (0.47)**
Zhang et al.'s SVM (Single sequence)[a]	20 × 7 + 3	74.8	49.1	67.9	0.41
Zhang et al.'s SVM (Multiple alignment)[a]	20 × 7 + 3	77.3	53.1	67.0	0.45
BetaTPred2[b]	–	75.5	49.8	72.3	0.43
Chou–Fasman[c]	–	74.9	46.1	16.9	0.16
1–4 and 2–3 correlation model[c]	–	63.2	35.3	60.4	0.21
Sequence coupled model[c]	–	50.6	31.7	88.4	0.23
Fuchs and Alix's[d]	–	74.8	48.8	69.9	0.42
Pham et al.'s SVM[e]	–	79.8	59.2	58.0	0.45

M is the number of input parameters.

[a]From (Zhang et al., 2005); [b]from (Kaur and Raghava, 2003); [c]from (Kaur and Raghava, 2002); [d]from (Fuchs and Alix, 2005); [e]from (Pham et al., 2005).

PSI denotes predicted secondary structure information in proteins by using of the PSIPRED; bold font denotes the most accurately predicted result. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

Table 3. The Predictive Results of Our Methods for the β-Turns in the 543 Proteins Using the 7-Fold Cross-Validation.

Method (parameter)	M	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	Mcc
PCSF (A_p)	20 × 7	75.0	44.3	38.2	0.23
ID (D_0)	400	60.0	39.0	63.4	0.28
ID (C_0)	6 × 7	59.1	36.6	58.9	0.25
ID (H_2)	36	61.3	35.6	55.2	0.24
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2))	4 × 2	74.4	46.3	59.4	0.41
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2) + PSI)	**4 × 2 + 3**	**76.6 (76.1)**	**47.6 (46.9)**	**70.2 (70.2)**	**0.43 (0.42)**

PSI denotes the predicted secondary structure information in proteins by using of the PSIPRED; bold font denotes the most accurately predicted result. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

PCSF method. However, Q_{obs} values are improved from about 32% of the PCSF algorithm to about 68% of the ID algorithm.

The Predictive Results by Using SVM

We train the classifiers with the LIBSVM program.[49] The radial basis function is selected as the kernel function; the optimized parameters are $c = 128$, $g = 0.5$, respectively, and weight factor is $w = 3$.

The results indicate that the Q_{pred} value is higher than the Q_{obs} value in the PCFS method, but the Q_{pred} value is lower than the Q_{obs} value in the ID method. Therefore, to enhance the prediction performance, the composite vectors with the above calculated PCSF and ID values are selected as the input parameters of the SVM. Better results are obtained in Table 2. The Mcc value is raised to 0.45 and overall prediction accuracy is increased to 77.3%.

Widely believed, the β-turn prediction accuracy can be greatly improved by using the secondary structure information.[13,21] Therefore, to further improve the predictive effect, the predictive secondary structure information is obtained by using the PSIPRED,[2] added in the input parameter. The predicted secondary structure of each residue is represented as: helix → (1, 0, 0), sheet → (0, 1, 0), and coil → (0, 0, 1). The PCSF, ID values, and predictive secondary structure information, together, are selected as input parameters. The performance is also shown in Table 2. The results indicate that the Q_{pred} value is 55.6% and the Q_{obs} value is 68.9%; Mcc is 0.47 and prediction accuracy is increased to 79.8%. This result is the highest achieved so far for predicting β-turn (bold font in Table 2). When the predicted secondary structure information is used in input the parameter, the prediction accuracy again gained about 3%.

Our method gives better rates than previous methods. The possible reasons are: first, the SVM is an extremely successful learning theory that usually outperforms other machine learning technologies such as artificial neural networks and nearest neighbor methods; and second, a new composite vector with ID, position conservation scoring function (PCSF) is employed. The ID and PCSF both algorithms may be extracted structure information of sequence. SVM with composite vector obtained Mcc value is raised to 0.45. The third reason is, the predicted secondary structure information by PSIPRED is used. The Mcc value is raised from 0.45 to 0.47. Comparing with previous Zhang et al.'s[21] and Pham et al.'s[27] SVM methods for the prediction of β-turn, using the improved composite vector in our method is a key step.

Some of the protein chains in our dataset may be used to train PSIPRED. To cross-validate the results, we have excluded those proteins from the nonredundant database of PSIPRED. As shown in Table 2; the difference in prediction performance is negligible.

Prediction of β-Turns in the 543 and 819 Chains Dataset

To evaluate the predictive method, the β-turns in the 543 and 819 chains dataset are predicted by using our method (results

Table 4. The Predictive Results of Our Methods for the β-Turns in the 819 Proteins Using the 7-Fold Cross-Validation.

Method (parameter)	M	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	Mcc
PCSF (A_p)	20 × 7	72.2	49.6	36.7	0.25
ID (D_0)	400	60.0	30.0	64.8	0.24
ID (C_0)	6 × 7	59.8	36.7	61.6	0.26
ID (H_2)	36	62.8	39.3	57.2	0.26
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2))	4 × 2	74.5	51.2	59.9	0.42
SVM (PCSF (A_p) + ID (C_0) + ID (D_0) + ID (H_2) + PSI)	**4 × 2 + 3**	**76.8 (76.5)**	**53.0 (52.6)**	**72.3 (72.4)**	**0.45 (0.45)**

PSI denotes the predicted secondary structure information in proteins by using of the PSIPRED; bold font denotes the most accurately predicted result. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

Table 5. The Predictive Results of Different Methods for the γ-Turns in the 320 Proteins Using the 5-Fold Cross-Validation.

Method (parameter)	*M*	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	Mcc
PCSF (A_p)	20 × 5	84.3	6.9	29.1	0.08
ID (A_0)	20 × 5	61.3	5.4	62.8	0.09
ID (H_0)	36	54.2	4.5	61.6	0.07
SVM (PCSF(A_p) + ID (A_0) + ID (H_0))	3 × 2	56.2	5.5	73.4	0.12
SVM (PCSF (A_p) + ID (A_0) + ID (H_0) + PSI)	**3 × 2 + 3**	**61.0 (60.8)**	**6.8 (6.3)**	**91.4 (90.0)**	**0.18 (0.17)**
SNNS[a,#]		74.0	6.3	83.2	0.17
Weka-logistic regression[a,#]		62.6	5.6	65.1	0.12
Weka-naive Bayes[a,#]		57.4	5.0	65.4	0.11
Weka-J48 classifier[a,#]		92.6	5.0	7.2	0.03
Sequence coupled model[a,*]		57.8	5.9	43.2	0.08
GOR[a,*]		75.5	6.1	45.5	0.09
Pham et al.'s SVM[b,#]		79.9	7.7	47.5	0.13

[a]From (Kaur and Raghava, 2003); [b]from (Pham et al., 2005).

*Using a single amino acid sequence and secondary structure information that from the PSIPRED.

#Using multiple alignment and secondary structure information; bold font denotes the most accurately predicted result. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

shown in Tables 3 and 4). The SVM's optimized parameters are $c = 2048$, $g = 0.5$, respectively, and weight factor is $w = 3$.

The results indicate that Q_{pred}, Q_{obs}, and Mcc values are respectively 47.6%, 70.2%, and 0.43, and prediction accuracy is 76.6% for 543 chains dataset. The results indicate that the Q_{pred} value is 53.0% and the Q_{obs} value is 72.3%; Mcc is 0.45 and prediction accuracy is 76.8% for 819 chains dataset. When the difference in performance of our method on three datasets was analyzed, it was observed that, although the performance decreased, the trend remained the same.

Prediction of γ-Turns in the 320 Chains Dataset

The above methods of predicting β-turn is also applied to the prediction of γ-turns in the 320 proteins with a sliding window of five amino acids (the PPMs contained 20 × 5 elements). The predictive results of the PSCF algorithm are shown in Table 5. The Mcc value only is 0.08. To compare our method with other methods, other method's results for using the same dataset are also shown in Table 5.

In the ID algorithm, A_0 and H_0 are, respectively, selected as the parameters of diversity source of the γ-turns and non-γ-turns. The performance is also shown in Table 5. The Mcc is 0.07−0.09, which is similar with 0.08 of the PCSF method.

Using the SVM method to predict γ-turns, the optimized parameters c and g are default; weight factor is $w = 29$. The above values of the PCSF (A_p), ID (A_0), and ID (H_0) are used to construct the composite vector as input parameters of SVM. The predictive results of SVM are shown in Table 5. The Mcc increases to 0.12.

To further improve the performance of the prediction, by adding the above composite vector and predictive secondary structure information to input parameter of SVM, better predictive results are obtained (see Table 5). The Mcc is increased to 0.18, better than the 0.13 in Pham et al.'s work[27] and the 0.17 in Kaur's[21] work. The Q_{obs} value is 91.4%, better than the 83.2% in Kaur's[21] work and the 47.5% in Pham et al.'s[27] work. Overall prediction accuracy is 61.0% (bold font in Table 5).

The above values are the most common measure of a method's overall performance; however, the Q_{total} can be misleading as γ-turn residues occur much less frequently than non-γ-

Table 6. The Predictive Results of Our Methods for the γ-Turns in the 346 Proteins Using the 5-Fold Cross-Validation.

Method (parameter)	*M*	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	Mcc
PCSF (A_p)	20 × 5	88.4	5.4	27.8	0.07
ID (A_0)	20 × 5	61.7	4.5	63.8	0.08
ID (H_0)	36	55.4	3.9	57.8	0.07
SVM (PCSF(A_p) + ID (A_0) + ID (H_0))	3 × 2	64.9	4.9	75.3	0.11
SVM (PCSF (A_p) + ID (A_0) + ID (H_0) + PSI)	**3 × 2 + 3**	**59.0 (58.9)**	**5.7 (5.6)**	**90.3 (89.2)**	**0.16 (0.16)**

PSI denotes the predicted secondary structure information in proteins by using of the PSIPRED; bold font denotes the most accurately predicted. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

Table 7. The Predictive Results of Our Methods for the γ-Turns in the 536 Proteins Using the 5-Fold Cross-Validation.

Method (parameter)	*M*	Q_{total} (%)	Q_{pred} (%)	Q_{obs} (%)	*Mcc*
PCSF (A_p)	20 × 5	75.9	6.1	38.2	0.08
ID (A_0)	20 × 5	59.0	5.2	63.9	0.09
ID (H_0)	36	56.4	4.6	58.8	0.08
SVM (PCSF(A_p) + ID (A_0) + ID (H_0))	3 × 2	58.4	5.3	73.4	0.12
SVM (PCSF (A_p) + ID (A_0) + ID (H_0) + PSI)	**3 × 2 + 3**	**58.5 (58.0)**	**6.8 (6.6)**	**92.7 (91.6)**	**0.18 (0.17)**

PSI denotes the predicted secondary structure information in proteins by using of the PSIPRED; bold font denotes the most accurately predicted result. Values shown in parentheses correspond to the results obtained by cross-validation of PSIPRED.

turn residues in proteins (1:~30). Therefore, one could easily achieve $Q_{total} \approx 97\%$ merely by predicting all residues to be non-γ-turn. For this reason, we consider Mcc, Q_{pred}, and Q_{obs} to be important indices.

The γ-turn results are obtained by using the same prediction method as with β-turns, only different parameters are selected, but given poor Q_{pred} and Mcc (only is 0.18) values in prediction results. This is definitely more unbalanced in the present dataset, which has a ratio of ~30:1 of non-γ-turn and γ-turn residues (β-turn (~3:1)). Moreover, a γ-turn consists of three residues and thus is much more flexible than a β-turn.[21]

Prediction of γ-Turns in the 346 and 536 Chains Dataset

To evaluate the predictive method, the γ-turns in the 346 and 536 chains dataset are predicted by using our method, and results are shown in Tables 6 and 7. The SVM's optimized parameters c and g are default; weight factor is $w = 30$ (for 346 chains) and $w = 26$ (for 536 chains), respectively.

The results indicate that the Q_{obs} value is 90.3%, the Mcc is 0.16 and the prediction accuracy is 59.0% for 346 chains dataset. The results indicate that the Q_{obs} value is 92.7%, the Mcc is 0.18, and the prediction accuracy is 58.5% for 536 chains dataset. The performance of three datasets is made by using the same rule.

Receiver Operating Characteristic Results

In addition, we calculated the area under the ROC curves of both SVM systems in the prediction β- and γ-turns. The performances of different systems have been evaluated by the ROC curves in Figure 2. For 320, 346, and 536 protein chain datasets, the corresponding areas under the ROC curves are 0.81, 0.78, and 0.79 for

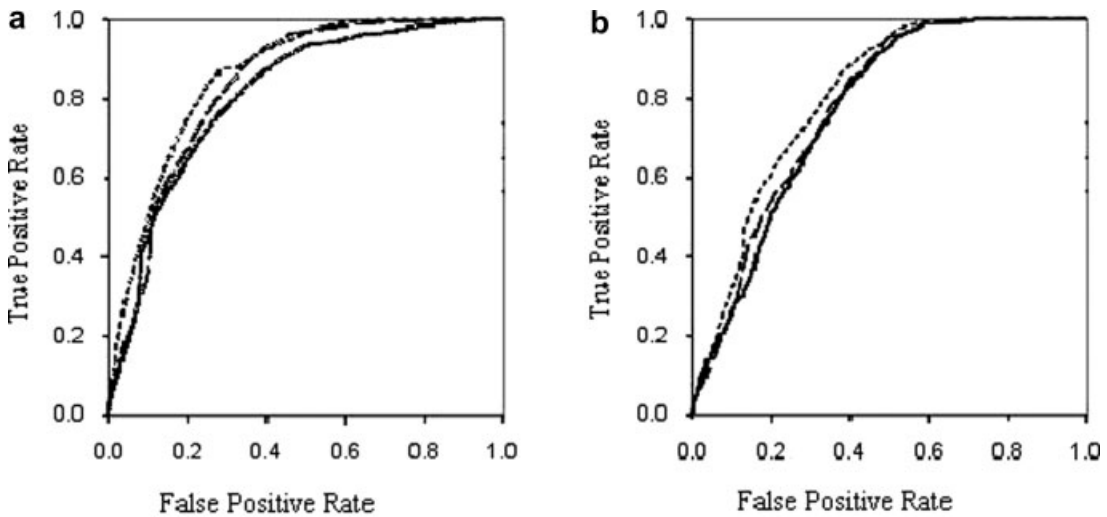

Figure 2. The ROC curves of both SVM systems of the prediction β-turns in 7-fold cross validation and γ-turns in 5-fold cross validation proteins. (a) shows ROC curves of β-turns in various dataset; Note: dotted line indicates a curve in the 426 protein chains dataset; solid line indicates a curve in the 543 protein chains; dashed line indicates a curve in the 819 protein chains. The corresponding areas under the ROC curves are 0.87, 0.82, and 0.84, respectively. (b) shows ROC curves of γ-turns in various dataset. Note: dotted line indicates a curve for the 320 protein chains; solid line indicates a curve for the 346 protein chains; dashed line indicates a curve in the 536 protein chains. The corresponding areas under the ROC curves are 0.81, 0.78, and 0.79, respectively.

γ-turns prediction, respectively. They are higher than 0.73 in Kaur and Raghava's method for γ-turns prediction.[33]

Conclusion

The above predicted results of the β- and γ-turns show that a single algorithm may provide partial information of a sequence. When the information of ID and PCSF are put together into the input parameters of SVM, the performance can be tremendously improved. It is possible that the SVM algorithm plays an information syncretizing role. In addition, the secondary structure information is also helpful to improve the predictive performance of the β- and γ-turns.

The successful prediction of β- and γ-turns in the proteins and, by using SVM with ID, PCSF values, and a predictive secondary structure as the input information, indicates a promising approach. Using them as SVM parameters can reduce dimension of input vector, improve calculating efficiency, and extract important classified information.

Acknowledgments

Authors are grateful to the editor and referees for their careful review and valuable comments on our manuscript.

References

1. Rost, B.; Sander, C. J Mol Biol 1993, 232, 584.
2. Jones, D. T. J Mol Biol 1999, 292, 195.
3. Petersen, T.N.; Lundegrad, C.; Neilsen, M.; Bohr, H.; Bohr, J.; Brunak, S.; Gippert, G. P.; Lund, O. Proteins 2000, 41, 17.
4. Cuff, J. A.; Barton, G. J. Proteins: Struct Funct Bioinform Genet 2000, 40, 502.
5. Guo, J.; Chen, H.; Sun, Z. R.; Lin, Y. Proteins: Struct Funct Bioinform 2004, 54, 738.
6. Wu, K. P.; Lin, H. N.; Chang, J. M.; Sung, T. Yi.; Hsu, W. L. Nucl Acids Res 2004, 32, 5059.
7. Rose, G. D.; Gierasch, L.; Smith, J. A. Adv Protein Chem 1985, 37, 1.
8. Takano, K.; Yamagata, Y.; Yutani, K. Biochemistry 2000, 39, 8655.
9. Cruz, X.; Thornton, J. M. Protein Sci 1999, 8, 750.
10. Rost, B.; Schneider, R.; Sander, C. J Mol Biol 1997, 270, 471.
11. Jones, D. T. Proteins Suppl 2001, 5, 127.
12. Chou, K. C. Anal Biochem 2000, 286, 1.
13. Kaur, H.; Raghava, G. P. Bioinformatics 2002, 18, 1508.
14. Kaur, H.; Raghava, G. P. Protein Sci 2003, 12, 627.
15. Chou, K. C.; Blinn, J. R. J Protein Chem 1997, 16, 575.
16. Shepherd, A. J.; Gorse, D.; Thornton, J. M. Protein Sci 1999, 8, 1045.
17. Chou, K. C. J Pept Res 1997, 49, 120.
18. Zhang, C. T.; Chou, K. C. Biopolymers 1997, 41, 673.
19. Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. J Pept Sci 2002, 8, 297.
20. Lin, T. H.; Wang, G. M.; Wang, Y. T. J Chem Inf Comput Sci 2002, 42, 123.
21. Zhang, Q. D.; Yoon, S.; Welsh, W. J. Bioinformatics 2005, 21, 2370.
22. Fuchs, P. F. J.; Alix, A. J. P. Proteins: Struct Funct Bioinform 2005, 59, 828.
23. Guruprasad, K.; Rajkumar, S. J Biosci 2000, 25, 143.
24. Chou, P. Y.; Fasman, G. D. Biochemistry 1974, 13, 211.
25. Gibrat, J.-F.; Garnier, J.; Robson, B. J Mol Biol 1987, 198, 425.
26. Wilmot, C. M.; Thornton, J. M. Protein Eng 1990, 3, 479.
27. Pham, T. H.; Satou, K.; Ho, T. B. J Bioinformatics Computat Biol 2005, 3, 343.
28. Street, T. O.; Fitzkee, N. C.; Perskie, L. L.; Rose G. D. Protein Science 2007, 16, 1720.
29. Bornot, A.; de Brevern, A. G. Bioinformation 2006, 1, 153.
30. Bystrov, V. F.; Portnova, S. L.; Tsetlin, V. I.; Ivanov, V. T.; Ochinnikov, Y. A. Tetrahedron 1969, 25, 493.
31. Milner-White, E. J.; Ross, B. M.; Ismail, R.; Belhadj-Mastefa, K.; Poet, R. J Mol Biol 1988, 204, 777.
32. Alkorta, I.; Suarez, M. L.; Herranz, R.; González-Muñiz, R.; García-López, M. T. J Mol Model 1996, 2, 16.
33. Kaur, H.; Raghava, G. P. S. Protein Sci 2003, 12, 923.
34. Guruprasad, K.; Shukla, S.; Adindla, S.; Guruprasad, L. J Peptide Res 2003, 61, 243.
35. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
36. Hutchinson, E. G.; Thornton, J. M. Protein Sci 1996, 5, 212.
37. Wasserman, W. W.; Sandelin, A. Nat Rev Genet 2004, 5, 276.
38. Kielbasa, S. M.; Gonze, D.; Herzel, H. BMC Bioinformatics 2005, 6, 237.
39. Cartharius, K.; Frech, K.; Grote, K.; Klocke, B.; Haltmeier, M.; Klingenhoff, A.; Frisch, M.; Bayerlein, M.; Werner, T. Bioinformatics 2005, 21, 2933.
40. Quandt, K.; Frech, K.; Karas, H.; Wingender, E.; Werner, T. Nucl Acids Res 1995, 23, 4878.
41. Kel, A. E.; GoBling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O. V.; Wingender, E. Nucl Acids Res 2003, 31, 3576.
42. Laxton, R. R. J Theor Biol 1978, 71, 51.
43. Li, Q. Z.; Lu, Z. Q. J Theor Boil 2001, 213, 493.
44. Zhang, L. R.; Luo, L. F. Nucleic Acids Res 2003, 31, 6214.
45. Chen, Y. L.; Li, Q. Z. J Theor Biol 2007, 245, 775.
46. Roobaert, D.; Hulle, M. M. In Proceedings of IEEE Neural Networks for Singnal Processing Workshop, Wisconsin, 1999; pp 77.
47. Schmidt, M.; Grish, H. Proc ICASSP'96, Atlanta, 1996; pp 105.
48. Hua, S. J.; Sun, Z. R. J Mol Biol 2001, 308, 397.
49. Chang, C. C.; Lin, C. J. 2001. LIBSVM. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
50. Deleo, J.M. In Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, IEEE, Computer Society Press, College Park, MD, 1993, pp 318.
51. Pánek, J.; Eidhammer, I.; Aasland, R. Proteins: Struct Funct Bioinform 2005, 58, 923.