

A Two-stage Classifier for Protein β -turn Prediction Using Support Vector Machines

Hua-Sheng Chiu, Hsin-Nan Lin, Allan Lo, Ting-Yi Sung and *Wen-Lian Hsu*

Abstract— β -turns play an important role in protein structures not only because of their sheer abundance, which is estimated to be approximately 25% of all protein residues, but also because of their significance in high-order structures of proteins. In this study, we introduce a new method of β -turn prediction that uses a two-stage classification scheme and an integrated framework for input features. Ten-fold cross validation based on a benchmark dataset of 426 non-homologue protein chains is used to evaluate our method's performance. The experimental results demonstrate that it achieves substantial improvements over BetaTurn, the current best method. The prediction accuracy, Q_{total} and the Matthews correlation coefficient (MCC) of our approach are 79% and 0.47 respectively, compared to 77% and 0.45 respectively for BetaTurn.

Index Terms— β -turn prediction, classification, protein secondary structure prediction, support vector machines

I. INTRODUCTION

IN protein tertiary structure prediction, secondary structure information is frequently used as a key element. The secondary structures of proteins consist of α -helices, β -sheets, tight turns, bulges, and random coils. Helices and sheets are referred to as regular secondary structures, while the other three are classified as irregular secondary structures [1]. Tight turns can be further decomposed into α -turns, β -turns, γ -turns, δ -turns, and π -turns according to the number of residues involved. A β -turn comprises four consecutive residues and satisfies the following additional conformational constraints: the distance between $C_\alpha(i)$ and $C_\alpha(i+3)$ is less than 7 Å, where $C_\alpha(j)$ denotes the alpha-carbon of an amino acid residue; and the tetrapeptide chain does not form an α -helix [1], [2].

β -turns play an important role in protein structures not only because of their sheer abundance, which is estimated to be approximately 25% of all protein residues [3], but also because of their significance in high-order structures of proteins [4]. β -hairpins, a super-secondary structure element found in anti-parallel β -sheets, are connected by β -turns. The latter are

also responsible for the compact globular shape of proteins because of their ability to reverse the direction of a protein chain within several residues. Moreover, β -turns tend to be located on solvent-exposed surfaces, which makes them accessible in molecular recognition processes [2], [5]. It has also been demonstrated that the formation of β -turns is a key factor in the early stages of protein folding [4]. Therefore, given the biological significance of β -turns, it would be useful to develop computational methods that accurately identify them in a protein sequence. Enhancing β -turn prediction would have a direct impact on molecular recognition studies and the identification of important structural motifs, such as β -hairpins. It would also contribute indirectly to the overall prediction of protein tertiary structures by improving secondary structure prediction.

The β -turn prediction methods developed thus far can be divided into two categories: those based on pure statistical methods, and those based on machine-learning approaches. Statistical methods use a positional preference approach [6]-[10], whereby the residue propensities within the turn at positions i to $i+3$ are used to calculate the positional frequencies and conformational parameters. More recently, a "correlation coupling effect" based on a positional preference method has been proposed. It considers the correlations of pairings of the 1-4 and 2-3 residues in a β -turn [11]. The second category, based on machine-learning approaches, includes neural network approaches [12]-[14], a k -nearest neighbor method [15], and a recently developed method based on SVM [16]. Inclusion of secondary structure information and positional specific substitution matrices from multiple sequence alignment in neural networks has been shown to improve prediction performance [13], [14]. Furthermore, the SVM method achieves the best performance, measured by Matthews correlation coefficient (MCC) at 0.45.

Although the proposed approach is based on SVM [17], [18], it differs from existing methods in several respects. First, the architecture of the classification scheme is a hierarchical two-stage model that has been used successfully in predicting the secondary structures of proteins [19]-[22]. Our method predicts the presence of β -turns with an additional layer of classification for coil and non-coil structures. Second, we have incorporated combinations of amino acid composition, evolutionary information, and predicted secondary structure as training input. Previous studies have not used these three types of information in an integrated manner. We also attempt to delineate the relationship between different combinations of

Manuscript received January 10, 2006.

Wen-Lian Hsu is affiliated with the Bioinformatics Lab., Institute of Information Science, Academia Sinica, Nankang, Taipei, 115, Taiwan (corresponding author to provide phone: 886-2-27883799 ext. 1804; fax: 886-2-27824814; e-mail: hsu@iis.sinica.edu.tw).

{Hua-Sheng Chiu, Hsin-Nan Lin, Allan Lo, Ting-Yi Sung} are affiliated with the same Institute as the corresponding author. (e-mail: {huasheng, arith, allanlo, tsung} @iis.sinica.edu.tw).

input features and the prediction results. Third, a different secondary structure prediction system, HYPROSP II [23], is used in place of PSI-PRED [24], [25]. In general, HYPROSP II outperforms PSIPRED in sensitivity, particularly for random coil structures. Our approach achieves a remarkable improvement over the current best method, BetaTurn [16], which it outperforms in various statistical measures.

II. SYSTEM AND METHODS

A. Two-stage Classifier

As the β -turn is a type of tight turn among coil structures, an intuitive way to classify β -turns is to separate the coil structures from their counterparts and then separate β -turns from other coil structures. The prediction problem can be regarded as a multiple classification process, in which all residues are divided into three classes: β -turns, non- β -turns or non-coils. We have developed a two-stage classification system to predict β -turns.

Our proposed method consists of two stages that predict: 1) coil residues, and 2) β -turns among coil residues, respectively. Fig. 1 shows the main steps of the proposed method. The first stage performs C/ \sim C classification intuitively. However, the classification yields a relatively lower recall rate, as shown by the results in Table I (a), which could affect the performance of the second stage. Therefore, we use five tertiary classifiers, previously used in the prediction of protein secondary structure elements (SSEs) [21], to construct the coil classifier. The first three classifiers consist of three one-versus-the-rest (H/ \sim H, E/ \sim E, C/ \sim C) classifiers for the entire dataset, and three one-versus-one (E/C, C/H, H/E) binary classifiers for the dataset of residues classified as \sim H, \sim E and \sim C, respectively, by the one-versus-the-rest classifiers. Our first three tertiary classifiers, each of which comprises two cascaded binary classifiers, are defined by (H/ \sim H, E/C), (E/ \sim E, C/H) and (C/ \sim C, H/E), called SVM_TREE1, SVM_TREE2, and SVM_TREE3, respectively. Since we are only concerned with coil prediction in this stage, the H/E binary classifier of SVM_TREE3 is discarded and only the binary classifier (C/ \sim C) is used. Fig. 2 illustrates these tertiary classifiers, which are structured like decision trees. The other two tertiary classifiers are called SVM_MAX_D and SVM_VOTE. In SVM_MAX_D, a residue is assigned to either the coil or the non-coil class that corresponds with the largest positive distance to the optimal separating hyperplane among SVM_TREE1, SVM_TREE2, and SVM_TREE3. The SVM_VOTE classifier combines the results of SVM_TREE1, SVM_TREE2, and SVM_TREE3 using the following simple voting rule. A residue is predicted to be a coil if most of the tertiary classifiers categorize it as such; otherwise, it is predicted to be a non-coil. In the second stage, we construct a binary classifier (β / \sim β) to categorize the β -turns among the coils predicted in the first stage.

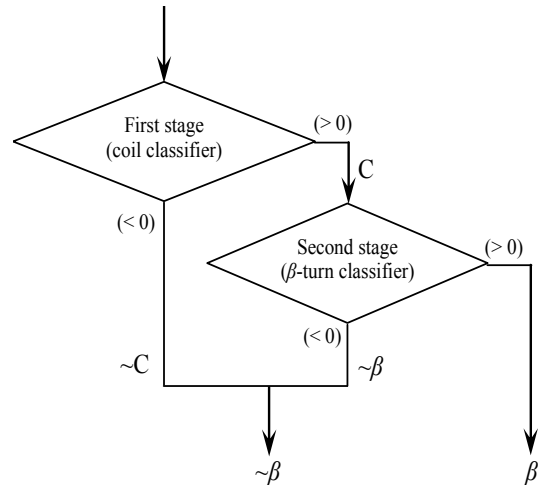


Fig. 1. The main steps of proposed two-stage classifier. Each residue of a query protein is classified as a non-coil if the output in the first stage is smaller than 0; otherwise, the second stage is used. If the output in the second stage is larger than 0, the residue is classified as a β -turn; otherwise a non- β -turn is assigned. Finally, each residue is predicted as either a β -turn (the β -turn residues in the second stage) or a non- β -turn (including the non-coil and non- β -turn residues in the first and second stages, respectively).

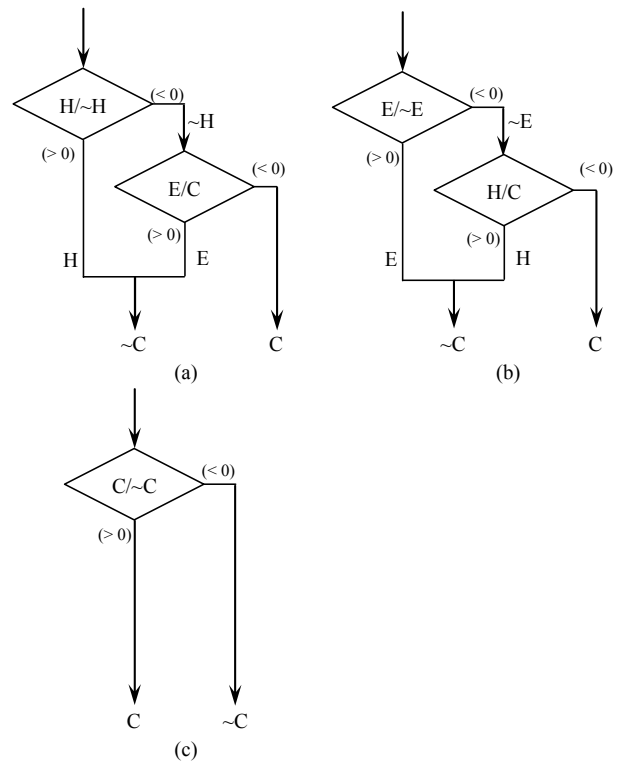


Fig. 2. The structures of the tertiary classifiers used in the first stage. (a) SVM_TREE1. (b) SVM_TREE2. (c) SVM_TREE3. Consider SVM_TREE1 as an example. A residue will be classified as a helix (H) if the output of the first binary classifier, H/ \sim H, is larger than 0; otherwise, the second binary classifier, E/C, is used. If the output of E/C is larger than 0, the residue is classified as a sheet (E); otherwise, a coil (C) is assigned. Each residue classified as a helix or a sheet is regarded as a non-coil; otherwise, it is a coil.

B. Feature Representation

The input features of our SVM classifiers are peptide sequences (i.e., amino acid compositions), position-specific scoring matrices (PSSM), and predicted SSE information. Given a protein of length n , we use a sliding window of size w to partition the protein into peptide sequences by shifting one residue at a time. Each peptide in turn is used to determine an input to the classifiers, which then predict the class of the residue located at the center of the peptide.

To capture the evolutionary information between the given protein and its remote homologues, we use PSI-BLAST [26] (the E-value threshold is set to 10^{-3} in three iterations) to search in the last-updated NCBI non-redundant database (nr) and construct a position-specific scoring matrix (PSSM) of size $n \times 20$. The $w \times 20$ submatrix of the PSSM corresponding to the underlying peptide is extracted and input to the classifiers in both stages.

In addition, we use amino acid (AA) composition information and the predicted SSE information of the underlying peptide as input features in the second stage. First, each residue of the peptide is represented by a unit vector of length 20 that has 1 in the position corresponding to the amino acid type of the residue, and 0 otherwise. SSE information is obtained by using HYPROSP II [23], a secondary structure prediction method. Therefore, the predicted secondary structure of each residue is represented as: helix \rightarrow (1,0,0), sheet \rightarrow (0,1,0), coil \rightarrow (0,0,1).

Let w_1 and w_2 be the size of the sliding windows used to partition peptides in the first and the second stages, respectively. The feature vectors of the classifiers in the first and second stages have lengths of $20 \times w_1$ and $(20 + 20 + 3) \times w_2$, respectively. Once the window exceeds the N- or C-terminal of the given protein, the “null” residue is represented by an all-zero vector. Taking account of the performance and time complexity, the window lengths w_1 and w_2 are optimized to 15 and 9, respectively. Furthermore, all attributes of the feature vectors are normalized within the range [0, 1].

C. Training and Testing

In both stages, we train the classifiers with the LIBSVM [27] program. The radial basis function (RBF) is selected as the kernel function, and the associated parameters (C, γ) are optimized with values of (1.8661, 0.1250). Since proteins are comprised of approximately 25% β -turns and 75% non- β -turns, using unbalanced classes for training could result in severe under-prediction (i.e., too many false negatives). Therefore, we set the cost weight of LIBSVM to 3 (75%/25%) for the negative samples to cope with the unbalanced data.

Ten-fold cross validation is used to evaluate our proposed method, whereby the dataset is first divided into ten subsets of equal size. Each subset in turn is tested using the classifiers trained on the remaining nine subsets. Since each residue of the whole dataset is only predicted once, the overall prediction accuracy is the percentage of correctly predicted residues.

D. Filtering

Our β -turn prediction method is applied to each residue in a query protein without reference to the prediction results of its neighboring residues. However, since β -turns have a length of at least four residues, we use the following “state-flipping” rules to correct our predictions [13]: 1) change an isolated non-turn prediction to a turn (i.e., t-t \rightarrow ttt); 2) change an isolated turn prediction to a non-turn (i.e., -t \rightarrow ---); 3) change an isolated pair of turn predictions to non-turns (i.e., -tt \rightarrow ----); and 4) change a non-turn prediction adjacent to three turn predictions to a turn (i.e., -ttt \rightarrow tttt- or -tttt). In our approach, the four rules are applied in the above sequence.

E. Performance Measures

In previous works, the following measures have usually been used to evaluate the performance of β -turn prediction: 1) Q_{total} (prediction accuracy), the percentage of correctly predicted residues; 2) Q_{pred} (probability of correct prediction, precision), the percentage of correct predictions among the residues predicted to be positives, where positives denote coils and β -turns in the first and the second stages, respectively; 3) Q_{obs} (sensitivity, recall), the percentage of observed (true) positives correctly predicted; and 4) MCC (the Matthews correlation coefficient). In this paper, we also use these measures and calculated them by the following equations:

$$Q_{\text{total}}^p (\%) = [(TP+TN)/(TP+TN+FP+FN)] \times 100 \quad (1)$$

$$Q_{\text{pred}}^p (\%) = [TP/(TP+FP)] \times 100 \quad (2)$$

$$Q_{\text{obs}}^p (\%) = [TP/(TP+FN)] \times 100 \quad (3)$$

$$MCC^p = \frac{[(TP)(TN)-(FP)(FN)]}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (4)$$

where p can be C (a coil in the first stage) or β (a β -turn in the second stage); TP denotes true positives; TN denotes true negatives; FP denotes false positives; and FN denotes false negatives. Since MCC takes account of both over- and under-predictions, it is the most robust and frequently used measure in β -turn prediction. If an approach has a higher MCC value in the range -1 to 1, its predictive ability is better.

III. RESULTS AND DISCUSSION

A. Dataset

We evaluate our method with a benchmark dataset consisting of 426 non-homologous protein chains that has been used in several related works [12]-[16]. The structure of each protein chain in the dataset is determined by X-ray crystallography at resolutions higher than 2.0 Å, and the sequence identity of any pair of protein chains is no more than 25%. We determine the location of observed β -turns in each chain by using the PROMOTIF program [28]. Each chain contains at least one β -turn, and some residues of a β -turn may be shared by other β -turns. In addition, the secondary structures of each protein chain are assigned by the popular DSSP program [29].

B. Experimental Results

The experimental results are reported in Table I. In the first stage, we use five tertiary classifiers, the performance of which is detailed in Table I (a). Among the first three classifiers, SVM_TREE1, and SVM_TREE2 outperform SVM_TREE3 by approximately 10% in most measures, particularly recall. The relatively poor performance of SVM_TREE3 provides evidence for not using C/~C classification directly in our first stage. However, SVM_TREE3 achieves the best precision among the five classifiers. The results of SVM_TREE1 and SVM_TREE2 are comparable, although the former has higher standard deviations in prediction accuracy of 0.11% and precision of 0.19%. Since SVM_MAX_D and SVM_VOTE combine the results of SVM_TREE1, SVM_TREE2, and SVM_TREE3, they perform better in most measures. However, SVM_VOTE is considered slightly better than SVM_MAX_D because only one performance measure of SVM_VOTE is lower than those of SVM_MAX_D. These conclusions agree with previous studies that used five tertiary classifiers for secondary structure prediction [19]-[22].

We also use AA composition, PSSM, and predicted SSE information as input features for the β -turn prediction in Stage 2 following the first-stage prediction results. The results are given in Table I (b). As mentioned earlier, SVM_MAX_D and SVM_VOTE achieve the best performance among the five classifiers with a prediction accuracy of approximately 79%,

recall of 48.11%, and MCC of 0.47. On closer examination, we observe the classifiers achieving the best performance in terms of a respective measure in coil prediction also achieve the best performance in the β -turn prediction. Thus, in our two-stage system, the performance of coil classifiers strongly influences the final results of β -turn prediction. The development of a more reliable and accurate coil classifier (e.g., SVM_MAX_D or SVM_VOTE) is therefore essential for β -turn prediction.

C. Effect of Different Combinations of Input Features on β -turn Prediction

In Stage 2 of β -turn prediction, we choose AA composition, PSSM, and predicted SSE information as the input features, and conduct experiments to examine the effects of different combinations of these features, as SVM_VOTE is used in Stage 1 for coil prediction. The results are detailed in Table II. In the baseline comparison, using PSSMs for prediction achieves a prediction accuracy of 74.72% and an MCC of 0.38. This outperforms prediction using AA composition, which yields a prediction accuracy of 72.12% and an MCC of 0.33. However, as coils are secondary structures, such SSE information may not always be available; therefore, we examine whether predicted SSE information can improve prediction. From Table II, we observe that the performance can be improved by incorporating such information. Specifically, the

TABLE I
EXPERIMENTAL RESULTS OF THE PROPOSED TWO-STAGE CLASSIFIERS

Coil classifiers	(a)				β -turn classifier	(b)			
	Q_{total}^C (%)	Q_{pred}^C (%)	Q_{obs}^C (%)	MCC ^c		Q_{total}^β (%)	Q_{pred}^β (%)	Q_{obs}^β (%)	MCC ^b
SVM_TREE1	77.27 (0.70)	71.12 (0.73)	81.71 (0.49)	0.55 (0.01)	AA + PSSM + SSE	78.35 (1.72)	66.44 (2.44)	47.81 (2.31)	0.46 (0.01)
SVM_TREE2	77.22 (0.59)	71.12 (0.54)	81.49 (0.54)	0.55 (0.01)		78.23 (1.65)	66.30 (2.45)	47.45 (2.56)	0.46 (0.01)
SVM_TREE3	76.84 (0.51)	74.90 (0.88)	71.54 (0.82)	0.53 (0.01)		77.54 (1.67)	69.34 (3.17)	46.33 (2.31)	0.45 (0.01)
SVM_MAX_D	77.50 (0.67)	71.31 (0.58)	82.08 (0.32)	0.56 (0.01)		78.99 (1.31)	66.75 (3.50)	48.11 (1.76)	0.47 (0.01)
SVM_VOTE	77.84 (0.55)	72.12 (0.54)	81.23 (0.27)	0.56 (0.01)		79.25 (1.10)	68.74 (2.34)	47.72 (1.65)	0.47 (0.01)

^aThe numbers in parentheses denote the standard deviation of the respective measure.

^bThe numbers in boldface denote the highest value of the respective measure.

TABLE II
THE PERFORMANCE OF DIFFERENT COMBINATIONS OF INPUT FEATURES

β -turn classifiers	Q_{total}^β (%)	Q_{pred}^β (%)	Q_{obs}^β (%)	MCC ^b
AA	72.12 (2.12)	59.56 (2.78)	39.44 (2.11)	0.33 (0.01)
PSSM	74.72 (1.80)	61.77 (2.11)	41.54 (1.81)	0.38 (0.01)
AA + SSE	75.85 (2.15)	65.22 (2.64)	45.32 (2.13)	0.43 (0.01)
PSSM + SSE	77.95 (1.42)	67.21 (2.45)	46.88 (1.81)	0.46 (0.01)
AA + PSSM + SSE	79.25 (1.10)	68.74 (2.34)	47.72 (1.65)	0.47 (0.01)

^aSVM_VOTE is used to classify the coil structures.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH AND BETA TURN

Methods	Q_{total}^β (%)	Q_{pred}^β (%)	Q_{obs}^β (%)	MCC ^b
SVM_VOTE /	77.95	67.21	46.88	0.46
PSSM + SSE	(1.42)	(2.45)	(1.81)	(0.01)
SVM_VOTE /	79.25	68.74	47.72	0.47
AA + PSSM + SSE	(1.10)	(2.34)	(1.65)	(0.01)
SVM-based (BetaTurn v1.1)	77.3	53.1	67.0	0.45

prediction accuracy and MCC are improved by at least 3% and 0.08, respectively. Since our experiments strongly support the integration of these three input features, we combine them and thereby obtain the best performance with a prediction accuracy of 79.25% and an MCC of 0.47.

D. Comparison with BetaTurn

Currently, the best β -turn prediction method is BetaTurn, an SVM-based approach that uses PSSM and predicted SSE information generated by PSI-PRED as input features to construct a classification model. We now compare our two-stage SVM-based method with BetaTurn. We use SVM_VOTE in Stage 1. Then, in Stage 2, we choose two sets of input features, one of which is the same as that used in BetaTurn and the other contains the three integrated input features discussed above. The results are summarized in Table III. Using the same features as BetaTurn, our proposed classifier in the first row obtains a higher MCC value (0.46) than BetaTurn (0.45). Both our models outperform BetaTurn in three performance measures, but not Q_{obs} . This superior performance can be attributed to the following factors: 1) we use a two-stage classifier, which reduces the classification complexity; and 2) we use a different protein secondary structure prediction system, HYPROSP II, which is more sensitive than PSI-PRED. Our study also shows that using a combination of AA composition, PSSM, and predicted SSE information as input improves prediction performance.

IV. CONCLUSIONS

The β -turn prediction approach proposed in this paper outperforms BetaTurn, the current best method. The prediction accuracy, Q_{total} , and the Matthews correlation coefficient (MCC) of our approach are 79% and 0.47 respectively, compared to 77% and 0.45 respectively for BetaTurn. These results demonstrate that our method is one of the best systems for the prediction of β -turns and non- β -turns. The improvement in the prediction accuracy of our system is attributable to two factors: 1) the use of a two-stage classification scheme; and 2) an integrated framework for feature input that comprises amino acid compositions, PSSMs, and predicted secondary structures. Our two-stage hierarchical classification scheme reduces the complexity of identifying β -turns by classifying random coil structures and thereby improves the prediction accuracy. Clearly, amalgamation of the input features in conjunction with effective filtering strategies results in enhanced performance. Our results suggest that other structure prediction systems, such as protein tertiary structures and β -hairpins, could be enhanced by incorporating the prediction results of our system.

Since our β -turn prediction method uses predicted SSE information as input, the development of more accurate SSE prediction systems would lead to incremental improvement in the prediction accuracy of our system. Finally, in our future work, we will incorporate additional features, such as solvent accessibility and hydrophobicity, into our approach. It would also be useful to construct a prediction system based on a

combination of most accurate methods.

REFERENCES

- [1] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Adv. Protein Chem.*, vol. 34, pp. 167-339, 1981.
- [2] G. D. Rose, L. M. and J. A. Smith, "Turns in peptides and proteins," *Adv. Protein Chem.*, vol. 37, pp. 100-109, 1985.
- [3] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-2637, 1983.
- [4] K. Takano, Y. Yamagata, and K. Yutani, "Role of amino acid residues at turns in the conformational stability and folding of human lysozyme," *Biochemistry*, vol. 39, pp. 8655-65, 2000.
- [5] K. C. Chou, "Prediction of tight turns and their types in proteins," *Anal. Biochem.*, vol. 286, pp. 1-16, 2000.
- [6] P. N. Lewis, F. A. Momany and H. A. Scheraga, "Chain reversals in proteins," *Biochem. Biophys. Acta.*, vol. 303, pp. 211-229, 1973.
- [7] P. Y. Chou and G. D. Fasman, "Conformational parameters for amino acids in helical, β -sheet and random coil regions calculated from proteins," *Biochemistry*, vol. 13, pp. 211-222, 1974.
- [8] C. M. Wilmot and J. M. Thornton, "Analysis and prediction of the different types of β -turns in proteins," *J. Mol. Biol.*, vol. 203, pp. 221-232, 1988.
- [9] E. G. Hutchinson and J. M. Thornton, "A revised set of potentials for beta-turn formation in proteins," *Protein Sci.*, vol. 3, pp. 2207-16, 1994.
- [10] F. J. Patrick and J. P. Alain, "High accuracy prediction of β -turns and their types using propensities and multiple alignments," *Proteins*, vol. 59, pp. 828-839, 2005.
- [11] C. T. Zhang and K. C. Chou, "Prediction of beta-turns in proteins by 1-4 & 2-3 correlation model," *Biopolymers*, vol. 41, pp. 673-702, 1997.
- [12] M. J. McGregor, T. P. Flores and M. J. E. Sternberg, "Prediction of β -turns in proteins using neural network," *Protein Eng.*, vol. 2, pp. 521-526, 1989.
- [13] A. J. Shepherd, D. Gorse, and J. M. Thornton, "Prediction of the location and type of beta-turns in proteins using neural networks," *Protein Sci.*, vol. 8, pp. 1045-55, 1999.
- [14] H. Kaur and G. P. Raghava, "An evaluation of beta-turn prediction methods," *Bioinformatics*, vol. 18, pp. 1508-14, 2002.
- [15] S. Kim, "Protein beta-turn prediction using nearest-neighbor method," *Bioinformatics*, vol. 20, pp. 40-4, 2004.
- [16] Q. Zhang, S. Yoon, and W. J. Welsh, "Improved method for predicting beta-turn using support vector machine," *Bioinformatics*, vol. 21, pp. 2370-4, 2005.
- [17] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273-293, 1995.
- [18] V. Vapnik, *Statistical Learning Theory*. New York: John Wiley and Sons, Inc., 1998.
- [19] L. H. Wang, J. Liu, Y. F. Li, and H. B. Zhou, "Predicting protein secondary structure by a support vector machine based on a new coding scheme," *Genome Inform Ser Workshop Genome Inform*, vol. 15, pp. 181-90, 2004.
- [20] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *J Mol Biol*, vol. 308, pp. 397-407, 2001.
- [21] H. Kim and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach," *Protein Eng.*, vol. 16, pp. 553-60, 2003.
- [22] M. N. Nguyen and J. C. Rajapakse, "Multi-class support vector machines for protein secondary structure prediction," *Genome Inform Ser Workshop Genome Inform*, vol. 14, pp. 218-27, 2003.
- [23] H. N. Lin, J. M. Chang, K. P. Wu, T. Y. Sung, and W. L. Hsu, "HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence," *Bioinformatics*, vol. 21, pp. 3227-33, 2005.
- [24] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J Mol Biol*, vol. 292, pp. 195-202, 1999.
- [25] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, pp. 404-5, 2000.
- [26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389-402, 1997.
- [27] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines". Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [28] E. G. Hutchinson and J. M. Thornton, "PROMOTIF--a program to identify and analyze structural motifs in proteins," *Protein Sci*, vol. 5, pp. 212-20, 1996.
- [29] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-637, 1983.