Using an Ensemble of Support Vector Machine Classifiers to Predict Protein Supersecondary Structural Motifs

Dongsheng Zou
College of Computer Science, Chongqing University, Chongqing 400044, China
Email: dszou@cqu.edu.cn

Zhongshi He and Yuan Yan
College of Computer Science, Chongqing University, Chongqing 400044, China
Email:{ zshe,yuany}@cqu.edu.cn

Abstract—**The success of human genome project and the rapid increase in the number of protein sequences entering into data bank have stimulated a challenging frontier: how to develop a fast and accurate method to predict the supersecondary structural motifs of protein. It could help to reduce the ever-widening gap between known sequences and unknown structure. To address this problem, a new method for prediction of protein supersecondary structural motifs is proposed in this paper. This method combines amino acid basic compositions with dipeptide components for feature representation of protein sequential patterns. An ensemble classifier based on Support vector machines is used to predict four kinds of supersecondary structural motifs in protein sequences. Total twenty-four increments of diversity are defined for each supersecondary structural motif. The method is trained and tested on ArchDB40 dataset containing 3088 proteins. The highest overall accuracy for the training dataset and the independent testing dataset are 74.8% and 69.3% respectively.**

Index-Terms—**supersecondary structural motifs; diversity measure; ensemble classifier; Support Vector Machines**

I. INTRODUCTION

Prediction of protein supersecondary structural motifs is a field which is being explored in bioinformatics. The information gained by predicting supersecondary structures may help to reduce the ever-widening gap between known sequences and unknown structure. In protein, supersecondary structures can be defined as two or several secondary structure units connected by polypeptide. The connective polypeptide, which does not contain any β-strand or α-helix residues, is called loop. According to the regular secondary structures connected by loops, supersecondary structures are divided into β–β (beta–loop–beta), β–α (beta–loop–alpha), α–β (alpha–loop–beta) and α–α (alpha–loop–alpha).

In the past few years, many methods are proposed for prediction of supersecondary structure [1-2]. Recently some attempts have been made to predict irregular secondary structures in a protein that includes α–turns [3-9]. Approaches for predicting supersecondary structural motifs fall into two main categories: finding different supersecondary structural motifs in a protein sequence and predicting special structural motifs. Sun [10] used an artificial neural network (ANN) to predict 11 different supersecondary motifs and achieved an accuracy ranging between 70% and 80% and a Matthew's correlation coefficient (MCC) between 0.40 and 0.50. Cruz [11] developed a method for predicting β-hairpins in a protein. They used a scoring scheme in which 14 scores were calculated on the basis of alignment and several properties, such as secondary structures, accessibility, specific pair interactions and non-specific distance. Using this approach they attained an accuracy of 47.7% (±3.9). Kuhn [12] attempted in 2004 to classify strand-loop -strand motifs by identifying local hairpins and non-local diverging turns using amino acid sequence as the input. This method achieved an accuracy of 77.3% (±6.1) by predicting the beginning and the ending of a hairpin and diverging turns. Kumar [13] used two machine-learning techniques, a support vector machine (SVM) and an ANN, both of which were based on several features, such as sequence information, evolutionary profile, surface accessibility and secondary structure information. They attained a highest accuracy of 79.2% in predicting β-hairpins. Hu [14] used SVM algorithm for β-hairpin prediction. They attained an overall accuracy of 79.9% and Matthew's Correlation coefficient of 0.59 on ArchDB40 dataset and an overall accuracy of 83.8% and Matthew's Correlation coefficient of 0.67 on Kumar's Dataset [13]. Hu [14] also used SVM algorithm for supersecondary structure prediction. They attained an overall accuracy of 71.7% and 64.5% on the training dataset and the testing dataset respectively. However, there is significant room for improvement of current approaches.

For all aforementioned methods, two key steps are extracting features from primary amino acid sequence and selecting suitable prediction engine. The performance of a method relies heavily on the sensitivity and selectivity of the corresponding feature vector and prediction algorithm. To satisfy the requirement of machine learning algorithm, each protein sequence need to be transformed into a fixed-length feature vector, some information of sequential feature would be lost,

doi:10.4304/jcp.6.10.2053-2059

especially the sequence-order effects. In this paper, one attempt to overcome this limitation is to combine different components of a feature vector for feature representation. Another particular challenge in training classifiers comes from the fact that the dataset used for training is unbalanced [15] to some extent. Standard machine learning algorithms without considering class-imbalance tend to be overwhelmed by the major class and ignore the minor one and lead to high false negative rate by predicting the positive point as the negative one[16]. To overcome this disadvantage, this paper tries to use an ensemble classifier to utilize all of the information extracted from the original sequences.

Against this background, in this paper, a novel method for prediction of supersecondary structural motifs s proposed. Our approach uses an ensemble classifier based on SVMs. The ensemble classifier can reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. In addition, this method combines amino acid basic compositions (AABC) with dipeptide components (DC) for feature representation of protein sequential patterns. Moreover, Increment of diversity (ID) is used for distance measure of two samples. ID is a kind of information description on state space and a measure of whole uncertainly and total information of a system [17]. The ID algorithm has been applied in the recognition of protein structural class [18] and the exon-intron splice site prediction [19]. By using AABC and DC as inputting parameters, four kinds of supersecondary structural motifs are predicted with ensemble classifier. We train and test our method on ArchDB40 dataset. The experimental results indicate that the proposed method could predict supersecondary structural motifs successfully.

II. Materials and Methods

A. ArchDB40 Dataset

ArchDB is based on DSSP [20] database and provided by Oliva [21], [22]. It contains the classification of protein loops from non redundant proteins of known structures (obtained from http:// sbi.imim.es/cgi-bin/archdb/loops.pl). In this paper, we use ArchDB40 subset that contains 3,088 non-redundant proteins with structures of resolution < 3 Å and < 40% sequence identity (ASTRAL SCOP 1.65).

The ArchDB40 subset contains 9,180 β–β motifs, 5,737 β–α motifs, 6,824 α–β motifs and 4,176 α–α motifs. According to Hu [14], the loop lengths of β-β motifs are mainly from 2 to 8 amino acid residues. Therefore, the supersecondary structural patterns with the loop length of 2-8 amino acids are extracted. There are 8,671 β-β motifs, 4,443 β–α motifs, 6,293 α–β motifs and 3,584 α–α motifs respectively. These patterns contain 94.5%, 77.4%, 92.2% and 85.8% of their total patterns respectively. The length of the supersecondary structural motif with 12 amino acids (2–8 amino acids in loop) is selected as the best fixed-length of pattern. It can cover more patterns and ensure minimum two consecutive amino acid residues in flanking loops.

B. Feature Representation

The concept of pseudo amino acid composition (PseAA), including the information of sequence order and length effects of protein, was introduced by Chou [23] and used widely to solve the problem of classification in bioinformatics fields [24-30]. In our method, AABC and DC are used to represent the compositional features of proteins.

Amino Acid Basic Composition

By calculating the number of twenty amino acids in L positions for all four kinds of supersecondary structural motifs, we can deduce four standard sources of diversity,

$$X^{\beta\beta} : \{N_{ij}^{\beta\beta} \mid i = 1, 2, \cdots, L;\ j = 1, 2, \cdots, 20\},$$

$$X^{\beta\alpha} : \{N_{ij}^{\beta\alpha} \mid i = 1, 2, \cdots, L;\ j = 1, 2, \cdots, 20\},$$

$$X^{\alpha\beta} : \{N_{ij}^{\alpha\beta} \mid i = 1,\ 2, \cdots, L;\ j = 1, 2, \cdots, 20\},$$

$$X^{\alpha\alpha} : \{N_{ij}^{\alpha\alpha} \mid i = 1, 2, \cdots, L;\ j = 1, 2, \cdots, 20\}$$

for β–β, β–α, α–β and α–α motif respectively. Here N_{ij} denotes the number of j_{th} amino acid occurring in the i_{th} position in the sequence.

Four standard measures of diversity corresponding to four standard sources of diversity can be deduced with similar equations as (1), namely

$$D(\xi) = M \log M - \sum_{ij} N_{ij}^{\xi} \log N_{ij}^{\xi}$$
$$(\xi = \beta\beta, \beta\alpha, \alpha\beta, \alpha\alpha)$$
$$M = \sum_{ij} N_{ij}^{\xi} \qquad (1)$$

Let $\log(0) = 0$ if N_{ij}^{ξ} equals zero.

A supersecondary structural motif S is also defined as a source of diversity in the same category space as $X^{\beta\beta}$, $X^{\beta\alpha}$, $X^{\alpha\beta}$ or $X^{\alpha\alpha}$. The measure of diversity of S can be calculated from the source of diversity. The increment of diversity for four sources of diversity S and $X^{\beta\beta}$, $X^{\beta\alpha}$, $X^{\alpha\beta}$ or $X^{\alpha\alpha}$ is

$$ID(X^{\xi}, S) = D(X^{\xi} + S) - D(X^{\xi}) - D(S)$$
$$(\xi = \beta\beta, \beta\alpha, \alpha\beta, \alpha\alpha) \qquad (2)$$

where $D(X^{\xi} + S)$ denotes the diversity measure of the mixed diversity source $X^{\xi} + S$. $D(X^{\xi})$ and $D(S)$ are the diversity measures of the diversity sources X^{ξ} and S respectively.

If

$$ID(X^{\xi},S)=Min\{ID(X^{\beta\beta},S),ID(X^{\beta\alpha},S),\\ ID(X^{\alpha\beta},S),ID(X^{\alpha\alpha},S)\} \quad (3)$$

Where ξ can be β–β, β–α, α–β or α–α and the operator Min means taking the minimum value among those in the parentheses. The ξ in (3) will give the sequence class to which the supersecondary structural motif S should belong.

The measure and the increment of diversity described above (1) and (2) are defined based on the diversity source in the category space of $L\times 20$ dimensions, L positions and twenty amino acids. The ID will be denoted as $ID\{L\times 20\}$.

Polypeptide Composition

The order and coupling information between amino acids in a protein sequence can be represented with Polypeptide Composition. The n -peptide characteristic matrix for a given sequential pattern can be defined by.

$$\varphi_n=\left\{\begin{matrix} N(A\overbrace{A\cdots A}^{n-1}) & \cdots & N(A\overbrace{V\cdots V}^{n-1}) \\ \cdots & N(\alpha_1\alpha_2\cdots\alpha_j\cdots\alpha_{n-1}\alpha_n)\cdots & \\ N(V\overbrace{A\cdots A}^{n-1}) & \cdots & N(V\overbrace{V\cdots V}^{n-1}) \end{matrix}\right\}_{20\times 20^{n-1}} \quad (4)$$

where $N(\alpha_1\alpha_2\cdots\alpha_j\cdots\alpha_{n-1}\alpha_n)$ denotes the absolute frequency of amino acid substring $\alpha_1\alpha_2\cdots\alpha_j\cdots\alpha_{n-1}\alpha_n$ occurring in the sequence. Here $\alpha_j(j=1,2,\cdots n)$ is one of the twenty amino acids. It can be deduced that φ_n contains 20^n elements. The noise of matrix data will increase dramatically when n is increasing. For sequential patterns of protein, polypeptide composition can represent the structure information when n equals 2. It will be dipeptide components in that case. The dipeptide components are important parameters for protein structure. As a result, the ID will be denoted as $ID\{20\times 20\}$.

Feature vectors

The feature information used in supersecondary structural motif prediction is mainly extracted from two classes of diversity source. The first class is built from AABC. It describes the basic composition of amino acids in a protein. The second class is built from DC. It describes the pair wise correlation of sequential two amino acids. The two classes of IDs are denoted as $ID_{\xi}\{L\times 20\}$ and $ID_{\xi}\{20\times 20\}$, respectively. Here L denotes the length of sequence pattern.

In our method for supersecondary structural motif prediction, twenty-four feature variables are defined as twenty-four increments of diversity ID_1 to ID_{24}. They are listed in Table 1. For example, the first variable ID_1 is defined by two diversity sources. The first is a β–β motif to be predicted. The second is a standard diversity containing all β–β motifs in the training set. The quantity can be calculated by use of (1) where the diversity measure is deduced from (1)

Each β-β motif is characterized by a vector of twenty-four dimensions, corresponding to the twenty-four variables (ID_1 to ID_{24}) defined above. The vector values are computed for all the supersecondary structural motifs in the training set, which is divided into four groups- β–β motifs, β–α motifs, α–β motifs and α–α motifs. Given a supersecondary structural motif S, ensemble classifier is applied to classify it as a β–β motif, β–α motif, α–β motif or α–α motif.

Table I. Feature representation of supersecondary structural motifs

ID notation	ID type & dimension	Source of information	ID defined by two sources	
			First source	Second source
ID_1, ID_2, ID_3	$ID_{\xi}\{L\times 20\}$	AABC	S	$D^1{}_{\beta\beta}$, $D^2{}_{\beta\beta}$, $D^3{}_{\beta\beta}$
ID_4, ID_5, ID_6	$ID_{\xi}\{L\times 20\}$	AABC	S	$D^1{}_{\beta\alpha}$, $D^2{}_{\beta\alpha}$, $D^3{}_{\beta\alpha}$
ID_7, ID_8, ID_9	$ID_{\xi}\{L\times 20\}$	AABC	S	$D^1{}_{\alpha\beta}$, $D^2{}_{\alpha\beta}$, $D^3{}_{\alpha\beta}$
ID_{10}, ID_{11}, ID_{12}	$ID_{\xi}\{L\times 20\}$	AABC	S	$D^1{}_{\alpha\alpha}$, $D^2{}_{\alpha\alpha}$, $D^3{}_{\alpha\alpha}$
ID_{13}, ID_{14}, ID_{15}	$ID_{\xi}\{20\times 20\}$	DC	S	$D^1{}_{\beta\beta}$, $D^2{}_{\beta\beta}$, $D^3{}_{\beta\beta}$
ID_{16}, ID_{17}, ID_{18}	$ID_{\xi}\{20\times 20\}$	DC	S	$D^1{}_{\beta\alpha}$, $D^2{}_{\beta\alpha}$, $D^3{}_{\beta\alpha}$
ID_{19}, ID_{20}, ID_{21}	$ID_{\xi}\{20\times 20\}$	DC	S	$D^1{}_{\alpha\beta}$, $D^2{}_{\alpha\beta}$, $D^3{}_{\alpha\beta}$
ID_{22}, ID_{23}, ID_{24}	$ID_{\xi}\{20\times 20\}$	DC	S	$D^1{}_{\alpha\alpha}$, $D^2{}_{\alpha\alpha}$, $D^3{}_{\alpha\alpha}$

Each β-β motif is characterized by a vector of twenty-four dimensions, corresponding to the twenty-four variables (ID_1 to ID_{24}) defined above. The vector values are computed for all the supersecondary structural motifs in the training set, which is divided into four groups- β–β motifs, β–α motifs, α–β motifs and α–α motifs. Given a supersecondary structural motif S , ensemble classifier is applied to classify it as a β–β motif, β–α motif, α–β motif or α–α motif.

The ID ($ID_j, j=1,2,...,24$) averaged over four kinds of supersecondary structural motifs in training set is denoted by $x_{\beta\beta}$, $x_{\beta\alpha}$, $x_{\alpha\beta}$ and $x_{\alpha\alpha}$ respectively. The corresponding covariance in four kinds of supersecondary structural motifs is represented by 24×24 matrix $\Sigma_{\beta\beta}$, $\Sigma_{\beta\alpha}$, $\Sigma_{\alpha\beta}$ or $\Sigma_{\alpha\alpha}$ respectively.

C. Ensemble classifier

Support vector machine

Support vector machine (SVM)classifier, motivated by results of statistical learning theory[31][32], is one of the most effective machine learning algorithms for many complex binary classification problems .Given the training set $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\in(X\times Y)^l\}$ when the penalty factor C and kernel function $K(.,.)$ are selected properly, we can construct a function

$$g(x)=\sum_{i:x\in X_+}\alpha_i K(x,x_i)-\sum_{i:x\in X_-}\alpha_i K(x,x_i)+b \tag{5}$$

where the non-negative weights α_i and b are computed during training by solving a convex quadratic programming. In order to estimate the probability of an unlabeled input x belonging to the positive class, $P(y=1\mid x)$, we map the value $g(x)$ to the probability by (Platt, 1999)

$$\begin{aligned}\Pr(y=1\mid x)&=P_{A,B}(g(x))\\&=1/[1+\exp(A*g(x)+B)]\end{aligned} \tag{6}$$

Where A and B are then obtained by solving the optimization problem

$$\begin{aligned}&\underbrace{\min}_{z=(A,B)}=F(z)=-\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i))\\&s.t.\quad ti=\begin{cases}(N_++1)/(N_++2) & if\, y_i=+1,\\ 1/(N_-+2) & if\, y_i=-1,\end{cases}\\&pi=P_{A,B}(g(xi)),\quad i=1,2,\cdots l\end{aligned} \tag{7}$$

Where N_+ and N_-, respectively, represent the number of positive and negative points in training set. Then the label of the new input x is assigned to be positive if the posterior probability is greater than a threshold, otherwise negative, i.e.

$$f(x)=\begin{cases}1, & if\ \Pr(y=1\mid x)>threshold\\ -1. & otherwise\end{cases} \tag{8}$$

where 1 corresponds to positive class, whereas -1 corresponds to negative class.

An ensemble of SVM classifiers

An ensemble of SVM classifiers is a collection of SVM classifiers, each trained on a subset of the training set(obtained by sampling from the entire training points) in order to get better results [33]. The prediction of the ensemble of SVMs is computed from the prediction of the individual SVM classifier, that is, during classification, for a new unlabeled input x_{test} ,the j -th SVM classifier in the collection returns a probability $P_j(y=1\mid x_{test})$ of x_{test} belonging to the positive class, where $j=1,2,\cdots m$ and m is the number of SVM classifiers in the collection. The ensemble estimated probability, $P_{Ens}(y=1\mid x_{test})$, is obtained by

$$P_{Ens}(y=1\mid x_{test})=(1/m)\times\sum_{j=1}^{j=m}P_j(y=1\mid x_{test}) \tag{9}$$

The architecture of the ensemble of SVM classifiers is shown in Fig.1.

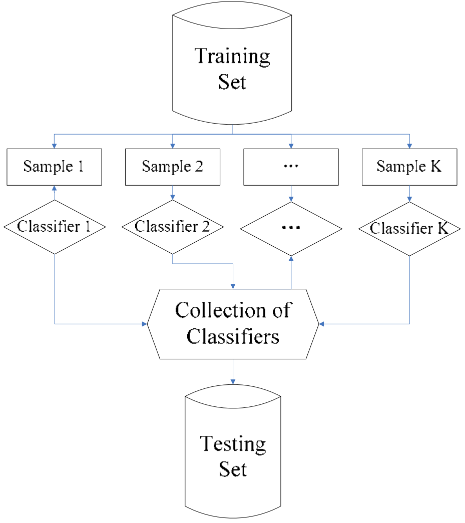

Fig.1. Architecture of the ensemble classifier fusing m SVM classifiers.

III. Results and Discussion

A. Performance Evaluation

The following standard measures are used to estimate the performance of our methods: (ⅰ) accuracy of prediction (S), (ⅱ) Matthew's correlation coefficients

(M_i), (iii) sensitivity (S_{ni}), (iv) specificity (S_{pi}).

$$S_{ni} = \frac{p_i}{p_i + u_i} \qquad S_{pi} = \frac{p_i}{p_i + o_i} \qquad S = \frac{\sum_i p_i}{N}$$

$$M_i = \frac{(p_i \times r_i) - (o_i \times u_i)}{\sqrt{(p_i + u_i)(p_i + o_i)(r_i + u_i)(r_i + o_i)}}$$

Here i denotes β–β, β–α, α–β and α–α supersecondary structures, p_i denotes the number of correctly predicted sequence segments for motif i, r_i denotes the number of segments that are correctly identified as something other than motif i, o_i denotes the number of segments which are not motif i but are predicted as motif i and u_i denotes the number of segments which are motif i but are predicted as something other than motif i. N denotes the sum of supersecondary structures.

B. Results

The ArchDB40 dataset is divided into training dataset and independent testing dataset in the same way with that of Hu [14]. Each kind of supersecondary structures is randomly divided into the training dataset and the independent testing dataset. The 7,000 out of 8,671 β–β motifs, the 3,500 out of 4,443 β–α motifs, the 5,000 out of 6,293 α–β motifs and the 3,000 out of 3,584 α–α motifs are selected as the training datasets. The remaining 1,671 β–β motifs, 943 β–α motifs, 1,293 α–β motifs and 584 α–α motifs are used as the independent testing datasets respectively.

For comparing the influence of different inputting parameters on the performance of our method, one of AABC and DC is firstly used as single inputting parameter and then AABC and DC are combined together as the compound inputting parameters.

The predictive results of the independent testing dataset and the training dataset are shown in Table 2. As shown in Table 2, accuracies of 72.0% and 67.2% are achieved for the training dataset and the testing dataset respectively, by using AABC as inputting parameters of our method. By using DC as inputting parameters, the accuracies are increased to 72.6% and 68.5%, respectively. By combining AABC with DC as inputting parameters, the accuracies are further increased to 74.8% and 69.3%, respectively. The results indicate that the order and coupling information in DC are very important for supersecondary structural motif prediction.

Compared with that of Hu [14], the highest accuracy of prediction is increased modestly by our method. For training dataset, accuracy increases from 71.7% to 74.8%; for testing dataset, accuracy increases from 64.5% to 69.3%. In addition, other measures such as S_{ni}, S_{pi} and M_i are also improved modestly. For example, sensitivity of β–β motifs increases from 78.6% to 82.1% for the training dataset when AABC and DC are combined as inputting parameters. Moreover, the performance measures for each kind of supersecondary structural motifs are relatively stable with our approach. However, these measures vary evidently with the method of Hu [14]. For example, the lowest sensitivity (S_{ni}) is 57.9% for α-α motifs while the highest one is 78.6% for β-β motif with Training dataset. The reason is that the sizes for each kind of supersecondary structural motifs are unbalanced. For example, the size of β-β motifs is 7000, which is much bigger than that of α-α motifs (3000) in Training dataset. As observed from Table2, we can summarize that this problem is partially solved with our approach by using an ensemble of SVM classifiers.

Generally, the results indicate the superior performance of our method over Hu's SVM [14].

Table II. Predictive results of ensemble classifier by using different inputting parameters

Parameter	Motifs	Training dataset				Testing dataset			
		S_n (%)	S_p (%)	M	S (%)	S_n (%)	S_p (%)	M	S (%)
HU's method	β-β	78.6	72.8	0.58	71.7	74.4	70.5	0.53	64.5
	β-α	67.5	69.2	0.57		53.4	60.0	0.52	
	α-β	73.4	70.7	0.54		68.9	67.9	0.55	
	α-α	57.9	65.8	0.55		51.2	62.4	0.52	
AABC	β-β	79.3	73.4	0.58	72.0	71.5	70.8	0.54	67.2
	β-α	78.1	76.0	0.58		64.1	65.3	0.53	
	α-β	75.9	75.1	0.57		65.6	66.9	0.53	
	α-α	70.5	72.7	0.56		63.0	62.2	0.52	
DC	β-β	79.9	73.8	0.58	72.6	70.6	71.0	0.54	68.5
	β-α	79.6	75.2	0.58		64.2	67.6	0.53	
	α-β	74.3	76.1	0.57		69.7	67.9	0.53	
	α-α	72.6	71.0	0.57		63.8	63.0	0.52	
AABC+DC	β-β	82.1	76.9	0.59	74.8	75.2	73.5	0.55	69.3
	β-α	79.9	75.3	0.58		67.4	66.6	0.53	
	α-β	77.3	74.1	0.57		70.5	69.2	0.54	
	α-α	75.5	72.6	0.57		64.9	68.3	0.53	

*The results of Hu's method are obtained from Ref. [14].

IV. Conclusion and Discussion

In this paper, we propose a new method for the prediction of supersecondary structural motifs. AABC and DC are combined to represent the features of protein sequential patterns and are vectored as the inputting parameters of an ensemble classifier. The experimental results show that the proposed method is quite suitable to predict supersecondary structural motifs. The advantage of ensemble classifier is that it can reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. Using the ID as inputting parameters can reduce dimension of inputting vector, improve calculating efficiency and extract more information for classification.

The better predictive results may be obtained by new approaches for feature representation of supersecondary structural motifs, such as taking the distribution of hydropathicity along protein sequence patterns into account. Other approaches may include finding new classifier to further improve the accuracy of prediction.

Acknowledgement

The authors thank the editor and referees for their careful review and valuable critical comments. This work is supported by the Fundamental Research Funds for the Central Universities of China No. CDJZR10180027.

References

[1] Fernandez-Fuentes N., Oliva B. and Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res., vol. 34, pp. 2085-2097, 2006.

[2] Chou K. C. Prediction of tight turns and their types in proteins. Anal. Biochem., vol. 286, no. 1, pp. 1-16, 2000.

[3] Kaur H. and Raghava G. P. S Prediction of Alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins, vol. 55, no. 1, pp. 83–90, 2004.

[4] Wang Y., Xue Z. D. and Xu J. Better prediction of the location of alpha-turns in proteins with support vector machine. Proteins: Struct. Funct. Bioinform., vol. 65, no. 1, pp. 49-54, 2006.

[5] Aurelie B. and Alexandre G. Protein beta-turn assignments. Bioinformation, vol. 1, no. 5, pp. 153-155, 2006.

[6] Kaur H. and Raghava G. P .S. A neural network method for prediction of β-turn types in proteins using evolutionary information. Bioinformatics, vol. 20, no. 16, pp. 2751–2758, 2004.

[7] Tho H. P., Satou K. and Tu B. H. Support vector machines for prediction and analysis of Beta and Gamma-turns in proteins. J. Bioinform. Comput. Biol., vol. 3, no. 2, pp. 343-358, 2005.

[8] Hu X. Z. and Li Q. Z. Using Support Vector Machine to Predict β- and γ-Turns in Proteins. J. Comput. Chem., vol. 29, no. 12, pp. 1867-1875, 2008.

[9] Jahandideh S., Sarvestani A. S., Abdolmaleki P., Jahandideh M. and Barfeie M. gamma-Turn types prediction in proteins using the support vector machines. J. Theor. Biol., vol. 249, no. 4, pp. 785-790, 2007.

[10] Sun Z., Rao X., Peng L. and Xu D. Prediction of protein super-secondary structures based on artificial neural network method. Protein Eng., vol. 10, no. 7 , pp. 763–769, 1997.

[11] Cruz X., Hutchinson E. G., Shepherd A. and Thornton J. M. Toward predicting protein topology: an approach to identifying Bhairpins. Proc. Natl. Acad. Sc. USA, vol. 99, no. 17, pp. 11157–11162, 2002.

[12] Kuhn M., Meiler J. and Baker D. Strand-loop- strand motifs: prediction of hairpins and diverging turns in proteins. Proteins, vol. 54, no. 2, pp. 282–288, 2004.

[13] Kumar M., Bhasin M., Navjot K. N.and Raghava G. P. S. BhairPred: Prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res., vol. 33, Web Server Issue, no. 1, pp. 154-159, 2005.

[14] Hu X. Z., Li Q. Z. Prediction of the β-hairpins in Proteins Using Support Vector Machine. Protein Journal, vol. 27, no. 1, pp. 115-122, 2008.

[15J apkowicz N. The class imbalance problem: significance and strategies. In: IC-AI'2000, Special Track on Inductive Learning Las Vegas, Nevada, 2000.

[16] Liu, X. Y., Zhou, Z. H.. The influence of class imbalance on cost-sensitive learning: an empirical study. In: Sixth IEEE International Conference on Data Mining (ICDM'06), Hong Kong, 2006.

[17] Laxton R. R. The measure of diversity. J. Theor. Biol., vol. 71, no. 1, pp. 51–67, 1978.

[18] Li Q. Z. and Lu Z. Q. The prediction of the structural class of protein: application of the measure of diversity. J. Theor. Biol., vol. 213, no. 3, pp. 493–502, 2001.

[19] Zhang L. R. and Luo L. F. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res., vol. 31, no. 21, pp. 6214–6220, 2003.

[20] Kabsch W. and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, vol. 22, no. 12, pp. 2577–2637, 1983.

[21] Oliva B., Bates P. A., Querol E., Aviles F. X. and Sternberg M. J. E.. An automatic classification of the structure of protein loops. J. Mol. Biol., vol.266, no. 4, pp.814–830, 1997.

[22] Espadaler J., Fuentes N. F., Hermoso A., Querol E., Aviles F. X., Sternberg M. J E., and Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res., vol. 32, Database Issue, pp. 185–188, 2004.

[23] Chou K. C. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins: Struct. Funct. Genet., vol. 43, no. 3, pp. 246-255, 2001.

[24] Lin H. and Li Q. Z. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Bioph. Res. Comm., vol. 354, no. 2, pp. 548–551, 2007.

[25] Chen Y. L. and Li Q. Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J. Theor. Biol., vol. 248, no. 2, pp. 377–381, 2007.

[26] Li F. M. and Li Q. Z.. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids, vol. 34, no. 1, pp. 119–125, 2008.

[27] Lin H.J. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition. J. Theor. Biol., vol. 252, no. 2, pp. 350-356, 2008.

[28] Lin H., Ding H., Guo F. B., Zhang A. Y. and Huang J. Predicting Subcellular Localization of Mycobacterial Proteins by Using Chou's Pseudo Amino Acid Composition. Protein Pept. Lett., vol. 15, no. 7, pp. 739-744, 2008.

[29] Lin H. and Li Q.Z. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Component. J. Comput. Chem., vol. 28, no. 9, pp. 1463-1466, 2007.

[30] Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, vol. 21, pp. 10 -19, 2005.

[31] Vapnik V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

[32] Vapnik V., 1998. Statistical Learning Theory. Wiley, New York.

[33] Dietterich, T.G. Ensemble methods in machine learning. In: Lecture Notes in Computer Science, vol. 1857, pp. 1–15, 2000.

Dongsheng Zou was born in Henan, P.R. China, in Sep 25, 1978. He received Ph.D. degree in computer science from Chongqing University, China in 2009. He is currently working in Chongqing University. His research interest includes data mining, computational biology and machine learning. He has published more than 10 papers in international academic journals including Journal of computational of Chemistry.

Zhongshi He was born in Sichuan, P.R. China, in Oct 17, 1965. He received Ph.D. degree in computer science from Chongqing University, China in 1996. He is currently a professor in Chongqing University. His research interest includes machine learning, computational biology and data mining. He has published more than 40 papers in international academic journals.

Yuan Yan was born in Chongqing, P.R. China, in Aug 18, 1986. She received bachelors degree in computer science from Chongqing University, China in 2010. Her research interest includesg, computational biology and machine learnin