ISSN : 0974 - 7435 Volume 10 Issue 12

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(12), 2014 [5986-5996]

Prediction of four kinds of supersecondary structures in enzymes by using svm based on scoring function

Sujuan Gao, Xiuzhen Hu*

College of Sciences, Inner Mongolia University of Technology, Huhhot, (P.R.CHINA)

E-mail: hxz@imut.edu.cn

ABSTRACT

Enzymes are a kind of protein that has catalytic function, the study of supersecondary structures in enzymes plays an important role in the structure and function of enzymes. Based on enzyme sequence information, four kinds of supersecondary structures in enzymes were researched for the first time. Amino acids of sites and dipeptide components of sites were selected as parameters, the predictive results were not ideal by using scoring function method; the better performance was obtained by using support vector machine (SVM), 40 scores for five selections of the best fixed-length pattern were selected as input parameters, the overall prediction accuracy in 7-fold cross-validation was 81.2%and the Matthews correlation coefficient was above 0.70. Therefore, SVM based on scoring function is an effective method to predict four kinds of supersecondary structures in enzymes.

KEYWORDS

Supersecondary structure; Scoring function; Support vector machine.

INTRODUCTION

Enzymes are also called biological catalysts, which produced by living cells with high specificity and catalytic efficiency. There need a series of chemical reactions to occur at the metabolism and almost all processes take place by catalytic efficiencies of the enzyme. That is to say, there will be no life phenomenon without enzyme. Therefore, study on enzyme structure and function is very important to the development of life science. In the past few years, much progress had been made towards the function of an enzyme, such as enzyme or non-enzyme[1-2], the classification of enzyme sub-classes[3-7]. However, studies on the structure of an enzyme were relatively limited.

Enzymes are a kind of protein that has catalytic function, they have the same primary sequence and advanced structure as common proteins. In protein, two or several secondary structure units are connected by loop (connective polypeptide without β-strand or α-helix) and formed a certain geometrical arrangement of the local space structure, it is called supersecondary structure or motif[8]. Simple supersecondary structures are classified into four types: β-loop-β, β-loop-α, α-loop-α and α-loop-β. Supersecondary structure is a building block of the tertiary structure of protein, it plays vital roles in a protein, such as providing folding stability, recognition and structure assembly. For these reasons, a number of supersecondary structure prediction methods have been developed in the past[9-17]. However, supersecondary structures in enzymes have their own characteristics, they often contain some binding sites and active sites, perform complicated biochemical function. For example, mitogen-activated protein kinase is an important transmitter signal from the cell surface to the nucleus, which contains a β-loop-α. β-strand at N-terminal and α-helix at C-terminal are separated by a deep channel, where is the binding site of ATP[18]. SnRK3 is also called calcineurin B-like calcium sensor-interacting protein kinase (CIPK), it has an inhibitory region in its C- terminus binding region, which combine with calcineurin B-like calcium sensor (CBL) to activate the kinase. CBL has a conservative core region contain four α-loop-α motifs, the conservative of every α-loop-α is relate to binding different kinase[19]. Hence, it is special significance in structure and function of enzyme that study on four kinds of supersecondary structures in enzymes.

In this paper, an attempt had been made to predict four kinds of supersecondary structures in enzymes. Supersecondary structures of 2261 enzymes, according to the regular secondary structures connected by loops, were divided into β-loop-β, β-loop-α, α-loop-α and α-loop-β. Based on amino acid sequence, the best fixed-length patterns contain 24 amino acids, which were generated using five methods: the first amino acid (beginning) of loop located the sixth position; end of loop located the nineteenth position; beginning of loop located the tenth position; end of loop located the fifteenth position; loop region located the center of pattern, amino acids of sites and dipeptide components of sites were selected as parameters, respectively, the lower accuracy were obtained by using scoring function method and SVM, amino acids of sites and dipeptide components of sites were together selected as parameters, the best result was obtained in support vector machine by using input parameters of 40 scores for five selections of the best fixed-length pattern.

MATERIALS AND METHODS

Materials

A dataset of 16,712 proteins with <95% sequence identity was downloaded from ASTRAL 1.75 of SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/), after excluding small proteins,14977 protein sequences were obtained. Blastcluster software was used for obtaining 8704 proteins with<25% sequence identity. Of which 4442 protein sequences were obtained whose length were more than100 residues with

resolution<3.0 Å. According to their main Enzyme Commission (EC) numbers[20], 2261 enzymes were obtained.

Secondary structure of each amino acid of all the proteins was assigned by using DSSP. The secondary structure 3-helix (G), 4-helix (H) and 5-helix (I) were expressed by α-helix; both the isolated β-bridge (B) and β-ladder (E) were expressed by β-strand; other secondary structure were expressed by loop, which contained S (bend), T (turn) and space. According to the regular secondary structures connected by loops, a total of 53367 unique supersecondary structure patterns were extracted, contain 14037 β-loop-β (EE),13391 β-loop-α (EH),13539 α-loop-α (HH)and 12400 α-loop-β (HE). The statistic analysis for supersecondary structure showed that loop lengths were mainly from 2 to 12 amino acids (Figure 1), there were 45506 motifs, contained 12956 EE, 10646 EH, 10682 HH and 11222HE, which took up 92.3%, 79.5%, 90.5% and 78.9% of the total number of corresponding motifs, respectively. Statistical analysis for sequence segment length in the motifs of loop length from 2 to 12 amino acids (Figure 2), the lengths were mainly from 6 to 30 amino acids, there were 41793 motifs, contained 12847 EE, 10090 EH, 8103 HH and 10753HE, which took up 99.2%, 94.8%, 75.8% and 95.8% of corresponding motifs, respectively. Thus we extracted the supersecondary structures of loop length from 2 to 12 amino acids and the sequence segment length from 6 to 30 amino acids as object of study.

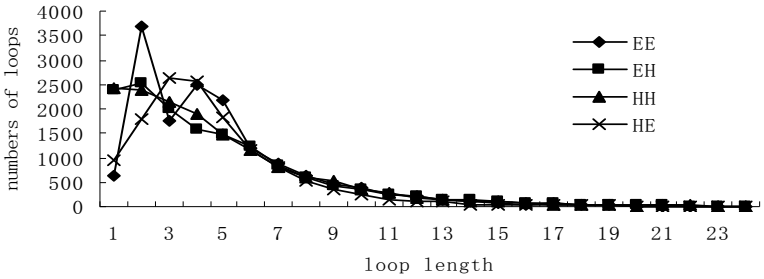

Figure 1 : The distribution of sequence numbers with different loop length in the supersecondary structures

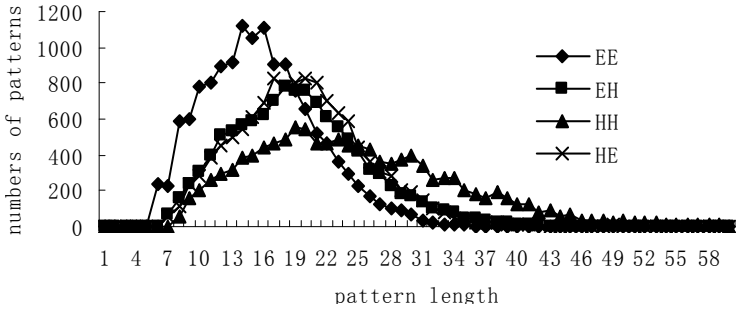

Figure 2 : The distribution of pattern numbers with different pattern length

METHODS

The selection of the best fixed-length pattern

Sequence segment lengths were statistically analyzed in the four kinds of supersecondary structures, the average lengths of EE, EH, HH, HE were respectively 15,19,24,19 amino acids, and the average length of α-helix was 9 amino acids, the average length of loop was 4 amino acids, the average length of β-strand was 5 amino acids. To ensure that the secondary structure connected by loops present completely in the fixed-length patterns, moreover, the N- and C- termini of loops in the motifs have relatively strong amino acid conservation, for example, glycine (G) present frequently at the both ends of loop[8]. Therefore, the fixed-length patterns of 24 amino acids were generated using five methods:

The first amino acid (beginning) of loop located the sixth position, end of loop located the nineteenth position, beginning of loop located the tenth position, end of loop located the fifteenth position and loop region located the center of pattern, examples shown in Figure 3.

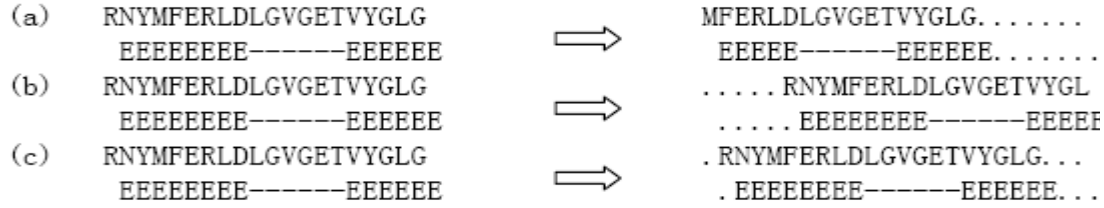

Figure 3 : The diagram of the best patterns fixed-length: (a) beginning of loop locates the sixth position (b) end of loop locates the nineteenth position (c) loop sequence locates the center. Note: first row is amino acid sequences, second row is secondary structures corresponding sequences, "."is a terminal residue.

STATISTICAL ANALYSIS OF THE POSITION CONSERVATION AND PARAMETER SELECTION

We performed a statistical analysis of the position conservation in the extracted segments using the WEBLOGO server (http:// weblogo.berkeley.edu/logo.cgi). Due to space constraints, sample of the position conservation was taken as example, shown in Figure 4.

In Figure 4(a), because the segments of EE were relatively shorter than other supersecondary structure types, they were gaps at positions 1 and 2, at positions from 3 to 15 and from 20 to 24, the most conservative amino acid was V, followed by Land A, at the positions 18 and19, the most conservative amino acid was G, followed by D and N; but in Figure 4(b), at positions from 1 to 16 and from 21 to 24, relatively conservative amino acids were L,A and E, at the position 18, the most conservative amino acid was L, at the position 19, the most conservative amino acid was D, followed by S and T, which indicated that the amino acids conservation was different at the end of loop in different supersecondary structures types. Comparing Figure 4(c) and Figure 4(d),in Figure 4(c), at positions from 1 to 12, relatively conservative amino acids were L,A and E, at positions from 16 to 24, the most conservative amino acids were V,L and A, at the position 15(end of loop), the most conservative amino acids were P and D, followed by G and A; but in Figure 4(d), at positions from 1 to 5, relatively conservative amino acids were L,A and E, at positions from 9 to 23,relatively conservative amino acids were V,L and A, at the position 24,the most conservative amino acids were S,F,T and V, at the position 6(beginning of loop), the most conservative amino acids were G and A, followed by L and K, which showed that the amino acids conservation was different at the both ends of loop in the same supersecondary structure. Comparing Figure 4(a) and Figure 4(c), for example, in Figure 4(a), at the position 19(end of loop), the most conservative amino acid was G, followed by D,N and P, in Figure 4(c), at the position 15(end of loop), the most conservative amino acid was P, followed by D,G and A, although the secondary structures connected by C-terminal both were β-strands, amino acid conservation was also different.

It is observed that different segment selection pattern have different amino acid conservation at the same position. Therefore, we selected amino acids of sites (20 amino acids and one terminal residue) as basic parameters, dipeptide components of sites were also selected as feature parameters of prediction in this paper.

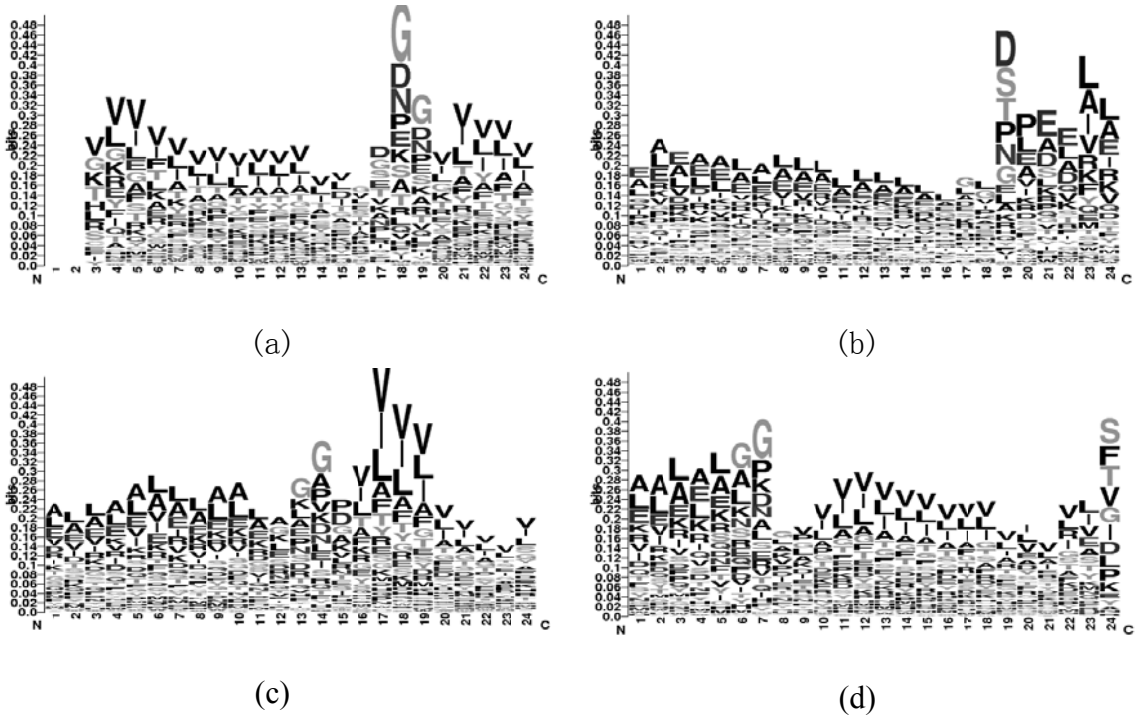

Figure 4 : Sample of the position conservation: (a) end of loop locates the nineteenth position of EE (b) end of loop locates the nineteenth position of HH (c) end of loop locates the fifteenth position of HE (d) beginning of loop locates the sixth position of HE. Note: the overall height of the stack indicates the position conservation, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position.

THE 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[21-22]. Amino acids of sites and dipeptide components of sites were selected as parameters, to consider the effect of position in supersecondary structure sequence segments, this paper used scoring function method to predict four kinds of supersecondary structures in enzymes.

(1) Position weight matrix (PWM)

The matrix element of position probability matrix represents probability at a corresponding position, which is defined as:

$$p_{ij} = \frac{n_{ij} + \sqrt{N_i}/l}{N_i + \sqrt{N_i}} \quad \textbf{(1)}$$

Here, amino acids of sites are selected as parameters, l equals 21, j denotes the 20 native amino acids and one terminal residue, n_{ij} denotes the real counts for amino acid j at the $i-th$ position of the sequence segments; when dipeptide components of sites are selected as parameters, l equals 441, j denotes the 441 dipeptides, n_{ij} denotes the real counts for dipeptide component j at the $i-th$ position of the sequence segments, N_i is the total number of the sequences.

Position weight matrix can be constructed according to position probability matrix, as following:

$$w_{ij} = \log \frac{p_{ij}}{p_{0j}} \quad \textbf{(2)}$$

Where p_{0j} is random probability, amino acids of sites are selected as parameters, the PWM includes $21 \times L$ elements; dipeptide components of sites are selected as parameters, the PWM includes $441 \times (L-1)$ elements, L is the length of the supersecondary structure sequence segments

(2) Conservation Index of Position

Conservation index vector of position in amino acid sequence reflects the difference of amino acid compositions between random sequences and one supersecondary structure sequences in the same position. The conservation index at the $i-th$ position may be defined by the following expressions[21]:

$$c_i = \frac{100}{\log l}(\sum_{j=1}^{l} p_{ij} \log p_{ij} + \log l) \quad \textbf{(3)}$$

$c_i \in [0,100]$, the higher C_i value is, the stronger conservation at the $i-th$ position is.

(3)Scoring Function

Scoring function can be calculated by following equation:

$$mss = \frac{\sum_{i=1}^{L} c_i (w_{ij} - w_{i,\min})}{\sum_{i=1}^{L} c_i (w_{i,\max} - w_{i,\min})} \quad \textbf{(4)}$$

$w_{i,\min}$ and $w_{i,\max}$ are the minimal and maximal values of position weight at the position i, respectively.

It is easily proved: $0 \le mss \le 1$, the value of mss shows that the degree of sequence close to known pattern sequences. Based on the probabilities of 21 amino acids (441dipeptide compositions) at the 24 positions of the patterns, the PWMs with 21×24 (441×23) elements for four kinds of supersecondary structure patterns can be constructed by using training datasets, the 4 scores are obtained for an arbitrarily sequence segment, supersecondary structure which the sequence segment should belong to may be predicted by the maximum among 4 scores.

SUPPORT VECTOR MACHINE

SVM is rigorously based on Vapnik's statistical learning theory[23-24], it has been widely used in protein structure prediction, protein subcellular location and protein folds classes[25-28]. SVM maps the input vector into a high dimensional feature space using a kernel function and to seek a separating hyperplane in this space, which make the distance among various samples achieve maximize, and realizes the maximize generalization ability. Common kernel function have the following four forms:

Linear function, Polynomial function, Radial basis kernel (RBF) function and Sigmoid kernel functions, here, we select RBF: $k(x, x_i) = \exp(-g\left\|x - x_i\right\|^2)$. In addition, SVM is a convex optimization problem, thus a local optimal solution is the global optimal solution. SVM has been compiled into the software packages, such as libsvm, mysvm, svmlight, and so on. In this article, we use the libsvm-2.91 software packages[29], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

According to the position conservation scoring function (PCSF) method (see section 2.2.3), the four kinds of supersecondary structures may obtain 4 standard position weight matrices by using the training dataset, every sequence segment in the testing datasets may obtain 20 scores for five selections of the best fixed-length pattern. Amino acids of sites and dipeptide components of sites are together selected as parameters, every sequence segment in the testing datasets may obtain 40 scores for five selections of the best fixed-length pattern. These scores can be used as feature parameters to input SVM.

K-FOLD CROSS-VALIDATION

We used the 7-fold cross-validation test to examine our prediction method. Four kinds of supersecondary structures were randomly divided into 7 subsets, the methods were trained on 6 subsets, and the performance was measured on the remaining seventh subset, this process was repeated 7 times so that each subset was tested.

PERFORMANCE EVALUATION

In order to evaluate the correct prediction rate and the reliability of a predictive method, the sensitivities (s_{ni}), specificities (s_{pi}), Matthew's correlation coefficients (M_i) and accuracies (S) are calculated by:

$$S_{ni} = [TP_i/(TP_i + FN_i)] \times 100\% \quad \textbf{(5)}$$

$$S_{pi} = [TP_i/(TP_i + FP_i)] \times 100\% \quad \textbf{(6)}$$

$$M_i = \frac{(TP_i \times TN_i) - (FN_i \times FP_i)}{\sqrt{(TP_i + TN_i) \times (TN_i + FP_i) \times (TP_i + FP_i) \times (TN_i + FN_i)}} \quad \textbf{(7)}$$

$$S = \frac{\sum_i TP_i}{N} \quad \textbf{(8)}$$

here i denotes four kinds of supersecondary structures (i=1 denotes EE, i=2 denotes EH, i=3 denotes HH, i=4 denotes HE), TP_i is the number of correctly predicted sequence segments for motif i, TN_i is the number of segments that are correctly identified as something other than motif i, FN_i is the number of segments which are not motif i but are predicted as motif i and FP_i is the number of segments of motif i that are missed by the prediction. N denotes sum of supersecondary structures.

RESULTS AND DISCUSSION

The predictive results by using pcsf algorithm

Amino acids of sites and dipeptide components of sites were, respectively, selected as parameters, by using five selections of the best fixed-length pattern, the predictive results by using PCSF algorithm in 7-fold cross-validation were shown in TABLE 1.

The performance indicate that individual predictive results of 21AA are good, such as the sensitivities of N6 and C19 in the EE motifs are 70.1% and 77.2%, respectively, however, the sensitivities of C19, N10, C15 and Center in the HH motifs are poor, in the HE motifs, the sensitivity of N6 is also only 25%; but the sensitivity of N6 in the HE motifs of 441JL is 79.2%, it is ideal, moreover, the specificity of C19 in the HH motifs is 81.9%, the sensitivity of N10 in the HE motifs is 81%, the sensitivities of C15 and Center in the EH motifs are79.7%and 80%, respectively, whereas the sensitivities of C19, N10, C15 and Center in the HH motifs are poor, too, which are slightly better than those of 21AA, in addition, the sensitivity of C19 in the EE motifs is only 22.3%. It can be found that the predictive results between 21AA and 441JL are complementary, therefore, the predictive effect should be improved if we combine two parameters. Because SVM algorithm can syncretize various parameters information, the calculated PCSF values will be selected as the input parameters of SVM.

TABLE 1 : The predictive results of PCSF algorithm using 7-fold cross-validation

	21AA					441JL				
	N6	**C19**	**N10**	**C15**	**Center**	**N6**	**C19**	**N10**	**C15**	**Center**
S_{n1}(%)	70.1	77.2	56.0	56.9	59.7	38.3	22.3	39.3	45.1	48.0
S_{n2}(%)	49.1	39.6	60.1	52.8	56.4	58.9	82.3	66.6	79.4	80.0
S_{n3}(%)	41.5	17.7	22.5	25.1	25.1	37.9	25.5	28.3	24.0	24.7
S_{n4}(%)	25.0	61.0	60.2	62.8	61.7	79.2	62.4	81.0	71.2	73.2
S_{p1}(%)	43.2	44.8	52.4	52.1	53.4	57.3	53.7	73.1	71.8	73.8
S_{p2}(%)	60.9	58.6	49.9	49.3	52.4	68.2	37.2	54.8	49.0	50.9
S_{p3}(%)	32.9	47.3	43.0	44.1	39.2	67.0	81.9	60.0	66.8	62.0
S_{p4}(%)	69.7	55.0	54.1	53.0	56.3	43.5	67.2	48.3	56.1	57.8
S (%)	46.7	50.4	51.0	50.6	52.0	54.3	49.3	55.0	56.5	58.0
M_1	0.23	0.30	0.29	0.29	0.32	0.27	0.16	0.38	0.41	0.45
M_2	0.31	0.27	0.28	0.25	0.29	0.45	0.24	0.37	0.39	0.41
M_3	0.09	0.14	0.13	0.15	0.12	0.37	0.36	0.27	0.29	0.27
M_4	0.26	0.33	0.32	0.32	0.35	0.32	0.44	0.38	0.41	0.45

Annotation: 21AA indicates that amino acids of sites are selected as parameters, 441JL indicates that dipeptide components of sites are selected as parameters. N6 indicates that beginning of loop locates the sixth position, C19 indicates that end of loop locates the nineteenth position, N10 indicates that beginning of loop locates the tenth position, C15 indicates that end of loop locates the fifteenth position and Center indicates that loop region locates the center of pattern.

THE PREDICTIVE RESULTS BY USING SVM

The above PCSF values are selected as the input parameters of SVM, by using 7-fold cross-validation, to further predict the four kinds of supersecondary structures in enzymes. The PCSF values of amino acids of sites are selected as the input parameters of SVM, the predictive results of every selection pattern are shown in TABLE 2. Comparing with the predictive results of the same parameter in the PCSF method (see TABLE1),the results are partial obviously improved, such as the sensitivities of N10,C15,Center in the EE motifs are improved from 56%,56.9%,59.7% of the PCSF algorithm to 78.9%,76.4%,81.5% of the SVM algorithm, respectively, but some predictive results are lower than

those of the PCSF method. The PCSF values of dipeptide components of sites are selected as the input parameters of SVM, the predictive results of every selection pattern are shown in TABLE 2. Comparing with the predictive results of the same parameter in the PCSF method (see TABLE1),the results are mostly obviously improved, such as the sensitivities of N6, C19, N10, C15, Center in the EE motifs are improved form 38.3%, 22.3%, 39.3%, 45.1%, 48% of the PCSF algorithm to 67.7%, 68.4%, 78.3%, 77.4%, 81% of the SVM algorithm, respectively, and so on, only individual predictive results are lower than those of the PCSF method.

In TABLE 2, amino acids of sites are selected as parameters, 20 PCSF values for five selections of the best fixed-length pattern can be input SVM (21AA{5}), the better prediction performance is obtained, comparing with the predictive results of every selection pattern before compositing (21AA), except that sensitivities of EE, EH, HE decrease slightly, such as the sensitivity of EE is 78.5% for 21AA{5}, but 78.9% for N10 of 21AA; dipeptide components of sites are selected as parameters, 20 PCSF values for five selections of the best fixed-length pattern can be input SVM (441JL{5}), all predictive results are superior to those of every selection pattern before compositing (441JL), the prediction accuracy can be tremendously improved, it is increased from maximum 63.8% to 78.2%, other predictive results are also above 73.6%; amino acids of sites and dipeptide components of sites are together selected as parameters, 40 PCSF values for five selections of the best fixed-length pattern can be input SVM(AAJL{5}), all measures are the best, such as the specificity of EH is raised to 84.2%, the specificity of HE is raised to 83%, the prediction accuracy is raised to 81.2%, Mcc is above 0.70 and so on. The results show that SVM algorithm can syncretize beneficial prediction information of four kinds of supersecondary structures in enzymes, it is an effective classifier.

TABLE 2 : The predictive results of SVM using 7-fold cross-validation

	21AA					441Jl					21AA {5}	441JL {5}	AAJL{5}
	N6	C19	N10	C15	Center	N6	C19	N10	C15	Center			
S_{n1}(%)	65.2	66.7	78.9	76.4	81.5	67.7	68.4	78.3	77.4	81.0	78.5	81.7	82.3
S_{n2}(%)	51.0	66.2	52.7	55.7	53.5	57.0	68.0	54.7	64.3	57.6	62.7	78.6	81.0
S_{n3}(%)	28.6	23.6	17.2	16.6	14.4	47.2	50.7	46.5	51.0	43.9	40.2	74.7	79.1
S_{n4}(%)	67.4	56.3	57.2	60.3	59.1	71.7	61.0	65.9	59.5	61.0	65.7	76.6	81.2
S_{p1}(%)	51.8	52.4	48.2	50.6	48.0	53.4	53.5	57.0	59.5	54.6	59.0	73.6	76.0
S_{p2}(%)	59.2	51.8	54.4	55.1	56.4	73.5	59.8	67.4	65.7	68.0	63.9	81.3	84.2
S_{p3}(%)	49.2	55.7	57.1	61.5	63.9	69.4	71.9	66.7	68.8	67.8	63.3	79.5	81.3
S_{p4}(%)	55.2	61.5	48.2	56.3	60.0	59.0	74.1	62.1	63.9	63.9	66.8	78.9	83.0
S(%)	54.5	54.8	53.3	54.2	54.1	61.8	62.8	62.2	63.8	61.9	63.0	78.2	81.2
M_1	0.34	0.35	0.36	0.37	0.37	0.39	0.40	0.48	0.50	0.47	0.50	0.68	0.70
M_2	0.34	0.34	0.30	0.32	0.33	0.50	0.45	0.45	0.49	0.47	0.46	0.72	0.76
M_3	0.21	0.22	0.19	0.21	0.20	0.45	0.49	0.43	0.47	0.42	0.38	0.70	0.74
M_4	0.38	0.38	0.37	0.35	0.38	0.46	0.53	0.46	0.45	0.45	0.50	0.69	0.75

Annotation: 21AA{5} indicated that combination of five selections of the best fixed-length pattern when amino acids of sites are selected as parameters, 441JL{5} indicated that combination of five selections of the best fixed-length pattern when dipeptide components of sites are selected as parameters, AAJL{5} indicates that combination of five selections of the best fixed-length pattern when amino acids of sites and dipeptide components of sites are together selected as parameters.

CONCLUSIONS

Supersecondary structures in enzymes play an important role in biochemical function of enzymes. In our work, four kinds of supersecondary structures in enzymes are theoretically predicted for the first time. A database containing 41793 motifs are constructed, the number of each kind motif has been greatly expanded, which can help validate the prediction of motif kinds. Based on enzyme sequence information, 24 amino acids are selected as the best fixed-length patterns by statistical analysis, according to the character of four kinds of supersecondary structures, five sequence segment selections are generated, amino acids of sites and dipeptide components of sites are selected as parameters, the 40 scores are obtained for an arbitrarily sequence segment by using scoring function method, the superior results are obtained in SVM by using input parameters of 40 scores.

The better prediction result obtained in this paper has following reasons: (1)The position conservation scoring function algorithm can extract important classified information and reduce dimension of input vector; (2) SVM algorithm can syncretize beneficial prediction information.

According to statistical analysis, amino acids of sites and dipeptide components of sites are selected as parameters in this article, other important features are not taken into account, such as hydrophobicity and flexibility of amino acid, which may be considered as prediction parameters in further work, it is hoped that the prediction accuracy is improved.

ACKNOWLEDGMENT

Authors are grateful to editor and referees for their careful review and valuable comments on our manuscript. This work was supported by the National Natural Science Foundation of China (30960090, 31260203).

REFERENCES

[1] Y.D.Cai, K.C.Chou; Using functional domain composition to predict enzyme family Classes, Journal of Proteome Research, **4**, 109-111 **(2005)**.

[2] Y.D.Cai, P.Z.Guo, K,C.Chou; Predicting enzyme family classes by hybridizing gene product composition and pseudo -amino acid composition, Journal of Theoretical Biology, **234**, 145-149 **(2005)**.

[3] K.C.Chou, Y.D.Cai; Using GO-PseAA predictor to predict enzyme sub-class, Biochemical and Biophysical Research Communications, **32**, 506-507 **(2004)**.

[4] H.B.Shen, K.C.Chou, EzyPred; A top-down approach for predicting enzyme functional classes and subclasses, Biochemical and Biophysical Research Communications, **364**, 53-59 **(2007)**.

[5] R.J.Shi, X.Z.Hu; Predicting enzyme subclasses by using support vector machine with composite vectors, Protein and Peptide Letters, **17(6)**, 599-604 **(2010)**.

[6] X.Z.Hu, Ting Wang; Prediction of enzyme subclass by using support vector machine based on improved parameters, 2011 Seventh International Conference on Natural Computation, 593-598 **(2011)**.

[7] Y.Wang, X.Z.Hu; Predicting of oxidoreductase and lyase subclasses by using support vector machine, 2011 10th IEEE/ACIS International Conference on Computer and Information Science, 27-31 **(2011)**.

[8] L.F.Yan Z.R.Sun; Moleculer structure of protein, Beijing, Tsinghua university Press, 43-56 **(1999)**.

[9] M.Kuhn, J.Meiler, D.Baker; Strand-loop-strand motifs, prediction of hairpin and diverging turns in proteins, Protein, **5**, 282-288 **(2004)**.

[10] X.Cruz, E.G.Hutchinson, A.Shepherd et al.; predicting protein topology: an approach to identifying Bhairpins, Proc.Natl.Acad.Sci. USA, **99**, ll157-1l162 **(2002)**.

[11] M.Kumar, M.Bhasin, N.K.Natt et al.; BhairPred, prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques, Nucleic Acids Research, **33**, 154-159 **(2005)**.

[12] X.Z.Hu, Q.Z.Li; The protein super-secondary structure recognition with the method of diversity measure, Acta Biophysica Sinica, **22(6)**, 424-428 **(2006)**.

[13] D.S.Zou, Z.S.He, J.Y.He et al; Supersecondary structure prediction using chou's pseudo amino acid composition, J.Comput.Chem., **32**, 271-278 **(2011)**.

[14] X.Z.Hu, Q.Z.Li; Prediction of the β-hairpins in proteins using support vector machine, The Protein Journal, **27(2)**, 115-122 **(2008)**.

[15] X.Z.Hu, Q.Z.Li, C.L.Wang; Recognition of β-hairpin motifs in proteins by using the composite vector, Amino Acids, **38(3)**, 915-921 **(2010)**.
[16] L.X.Sun, X.Z.Hu, S.B.Li; Predicting βαβ motifs based on SVM by using the ID and MS values, 2012 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012), 910-914 **(2012)**.
[17] Van Du T.Tran, Philippe Chassignet, Jean-Marc Steyaert; Supersecondary structure prediction of transmembrane beta-barrel proteins, Methods in molecular biology, **932**, 277-294 **(2013)**.
[18] Z.Wang, P.C.Harkins, R.J.Ulevitch, J.H.Han, M.H.Cobb, E.J.Goldsmith; The structure of mitogen-activated protein kinase p38 at 2.1-Å resolution, Proc.Natl.Acad.Sci. USA, **94(6)**, 2327-2332 **(1997)**.
[19] O.Batistic, J.Kudla; Integration and channeling of calcium signaling through the CBL calcium sensor/CIPK protein kinase network, Planta, **219(6)**, 915-924 **(2004)**.
[20] E.C.Webb, Enzyme Nomenclature, Academic Press, SanDiego, **(1992)**.
[21] K.Cartharius, K.Frech, K.Grote et al.; Mat inspector and beyond, Promoter analysis based on transcription factor binding sites, Bioinformatics, **21(13)**, 2933-2942 **(2005)**.
[22] A.E.Kel, E.GoBling I.Reuter et al.; Matchtm, A tool for searching transcription factor binding sites in DNA sequences, Nucleic.Acids.Research, **31(13)**, 3576-3579 **(2003)**.
[23] V.Vapnik; The nature of statistical learning theory, New York, Springer, **(1995)**.
[24] V.Vapnik; Statistical learning theory, Wiley-Interscience, **(1998)**.
[25] X.Z.Hu, Q.Z.L; Using support vector machine to predict β-turns and γ-turns in proteins, Computational Chemistry, **29(12)**, 1867-1875 **(2008)**.
[26] K.C.Chou, Y.D.Cai; Using functional domain composition and support vector machines for prediction of protein subcellular location, Journal of Biological Chemistry, **227**, 45765-45769 **(2002)**.
[27] C.H.Q.Ding, I.Dubchak; Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics, **17(4)**, 349-358 **(2001)**.
[28] J.Y.Shi, Z.Pan, S.W.Zhang, Y.Liang; Protein fold recognition with support vector machines fusion network, Progress in Biochemistry Biophysics, **3(2)**, 155-162 **(2006)**.
[29] C.C.Chang, C.J.Lin; Libsvm, A library for support vector machines, Software available at http: //www. Csie.ntu.edu.tw/cjlin/libsvm, **(2001)**.