Saudi Journal of Biological Sciences (2016) 23, 66–71

King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Prediction of complex super-secondary structure βαβ motifs based on combined features

Lixia Sun, Xiuzhen Hu *, Shaobo Li, Zhuo Jiang, Kun Li

College of Sciences, Inner Mongolia University of Technology, Hohhot 010051, China

Received 11 September 2015; revised 8 October 2015; accepted 12 October 2015
Available online 12 November 2015

KEYWORDS

Structure prediction;
Complex super-secondary structure;
Combined features;
Support Vector Machine

Abstract Prediction of a complex super-secondary structure is a key step in the study of tertiary structures of proteins. The strand-loop-helix-loop-strand (βαβ) motif is an important complex super-secondary structure in proteins. Many functional sites and active sites often occur in polypeptides of βαβ motifs. Therefore, the accurate prediction of βαβ motifs is very important to recognizing protein tertiary structure and the study of protein function. In this study, the βαβ motif dataset was first constructed using the DSSP package. A statistical analysis was then performed on βαβ motifs and non-βαβ motifs. The target motif was selected, and the length of the loop-α-loop varies from 10 to 26 amino acids. The ideal fixed-length pattern comprised 32 amino acids. A Support Vector Machine algorithm was developed for predicting βαβ motifs by using the sequence information, the predicted structure and function information to express the sequence feature. The overall predictive accuracy of 5-fold cross-validation and independent test was 81.7% and 76.7%, respectively. The Matthew's correlation coefficient of the 5-fold cross-validation and independent test are 0.63 and 0.53, respectively. Results demonstrate that the proposed method is an effective approach for predicting βαβ motifs and can be used for structure and function studies of proteins.

1. Introduction

In proteins, if two secondary structure units are connected by a polypeptide (loop) with a specific arrangement of geometry, the resulting structure is referred to as a super-secondary structure or motif. Two or more super-secondary structures further fold into a complex super-secondary structure (Kuhn et al., 2004). The βαβ motif is a complex super-secondary structure in proteins (Yan and Sun, 1999), and it often appears in *Bacillus subtilis* proteases (Blundell et al., 1987). In strand-loop-helix-loop-strand structures, if there are one or more hydrogen bonds between two parallel β-strands, the structure is referred to as a βαβ motif.

The structure of a protein determines its function (Sun et al., 1997; Chou and Zhang, 1995; Chou, 1995). Thus, the prediction of protein structure is quite important in function research. At present, it is difficult to directly predict the tertiary structure from a protein sequence. Moreover, the super-secondary structure is a bridge between the secondary

* Corresponding author. Tel.: +86 471 6576281; fax: +86 471 6575863.
E-mail address: hxz@imut.edu.cn (X. Hu).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.sjbs.2015.10.005

structure and tertiary structure, especially the complex super-secondary structure. Therefore, the prediction of complex super-secondary structure is a key step for the study of tertiary structure. With the increasing number of known protein structures and the well-developed feature selection algorithms such as mRMR (Peng et al., 2005), it is possible to develop theoretical methods to predict the complex super-secondary structure βαβ motifs in proteins.

The βαβ motif is an important complex super-secondary structure in proteins. In addition, many functional sites and active sites often occur in the polypeptides of βαβ motifs (Yan and Sun, 1999), including ADP-binding sites, FAD-binding sites, NAD-binding sites and other such functional sites (Wierenga et al., 1986). Therefore, the accurate prediction of βαβ motifs is very important to recognizing protein tertiary structure and the study of protein function.

Study of the βαβ motif began in 1983, where Taylor and Thornton correctly predicted the βαβ motifs with 70% accuracy in 16 α/β type proteins during predictions of a super-secondary structure (Taylor and Thornton, 1983). In 1984, they applied their method to identify 66 βαβ motifs in 18 proteins of α/β class with 75% accuracy (Taylor and Thornton, 1984). In 1986, Wierenga et al. predicted the occurrence of ADP-binding βαβ folds in the proteins from the PIR database, which contains 2676 proteins, by using amino acid sequence fingerprinting (Wierenga et al., 1986), but their dataset only involved the βαβ motif structure. Because the number of known protein structures was not sufficient at that time, they were limited to predicting the βαβ motif by using statistical methods. In more recent years, high-throughput technologies, information technology and computer technology have rapidly developed. The number of known protein structures has greatly increased, and it is feasible to predict the βαβ motif in a protein by using theoretical methods. In this study, we predicted the complex super-secondary structure βαβ motif, based on the principles and method which have been successfully used to predict β-hairpins of proteins in our previous work (Jia and Hu, 2011; Hu et al., 2010; Hu and Li, 2008).

The key step of complex super-secondary structure prediction is to construct a reasonable dataset and to select the best characteristic parameters and algorithm. In this paper, we constructed complex super-secondary structure βαβ motif datasets from protein structure data and used the sequence information, predicted structure information and function information to express the sequence features. To avoid an overfitting phenomenon when the higher dimension features are used in a Support Vector Machine algorithm, the dimensions of the amino acid component of position that is used to represent the sequence information were optimized by mRMR (Peng et al., 2005). Good predictive results were achieved according to 5-fold cross-validation and independent test.

2. Materials and methods

2.1. Materials

2.1.1. Dataset

In this paper, the βαβ motif datasets were constructed by using DSSP (Kabsch and Sander, 1983) in the following four steps:

(1) A dataset of 16,712 protein chains with < 95% sequence identity was downloaded from ASTRAL 1.75 of SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/). We then deleted small proteins, and 14,977 protein chains were obtained. The dataset contained four classes of proteins: all α proteins, all β proteins, α/β proteins, and α + β proteins.

(2) BLAST software (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.6/) was used for obtaining proteins with < 25% sequence identity. In total, 8704 protein chains were obtained. To accurately define the secondary structure in DSSP, 4442 protein chains were obtained whose length was more than 100 residues with a resolution < 3.0 Å.

(3) Secondary structure was assigned to each amino acid of all the protein chains using DSSP. DSSP defines 8 states: H (α-helices), G (3_{10}-helices), I (π-helices), B (single β-bridge), E (β-ladder), T (hydrogen bonded turn), S (bend) and blank. However, prediction methods are normally assessed for only 3 states (H, C, and E), and therefore, the 8 states have to be reduced to 3 states. The secondary structures G, H and I were expressed by H. Both B and E were expressed by E, and the remaining states were expressed by C. Overall, 11,736 ECHCE patterns were obtained.

(4) In ECHCE patterns, if there were one or more hydrogen bonds between two parallel β-strands, then the structure was designated as a βαβ motif. Totally, there were 1635 protein chains which contained at least one βαβ motif. Overall, 4277 βαβ motifs and 3366 non-βαβ motifs were obtained.

2.1.2. The statistical analysis of the sequence segments

Loop-helix-loop (loop-α-loop) is a nucleation structure in the βαβ (β-loop-α-loop-β) motif. To choose the study objects, we performed a statistical analysis on the loop-α-loop structure. Results of this analysis are shown in Fig. 1.

Fig. 1 shows that the loop-α-loop lengths of the βαβ motifs and non-βαβ motifs are mainly concentrated between 10 and 26 amino acids which accounted for 85.6% of the total motifs. Therefore, we extracted the loop-α-loop that was 10–26 amino acids in length to study (Taylor and Thornton, 1984).

Furthermore, the statistical analysis of the segment length is performed and the results are shown in Table 1. The ideal fixed-length pattern was selected as 32 amino acids based on the average length of βαβ motifs and non-βαβ motifs, and Kumar's segment selection method in the prediction of β-hairpins (Kumar et al., 2005) (see the following section).

2.1.3. The selection of the fixed-length pattern

According to the statistical analysis of the loop-α-loop structure and the principles of Kuhn et al.(2004) and Hu and Li, 2008 used in β-hairpin prediction, two selection methods were generated. In the first method the beginning of the left loop was considered to be located at the fifth position. In the second method the end of the right loop was considered to be located at the twenty-eighth position in a fixed-length pattern. The two methods ensured that all loops can be included in the fixed-length pattern.

Based on the central alignment principles of Kumar's method in β-hairpin prediction (Kumar et al., 2005), the

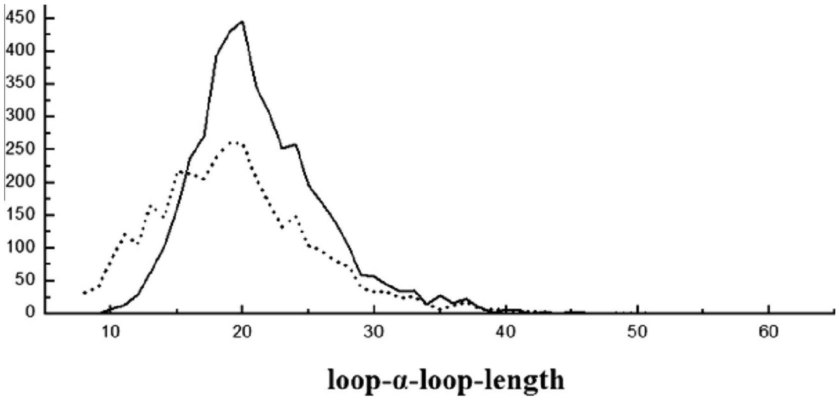

Figure 1 The distribution of the length of the loop-α-loop motif. The solid line is used to represent βαβ motifs, and dotted line is used to represent the non-βαβ motif.

Table 1 The length of the patterns in the datasets.

Motifs	βαβ	Non-βαβ
Longest length	65	66
Shortest length	15	12
Average length	31.6	29.3

loop-α-loop should be located in the center of the fixed-length pattern, the number of amino acids of the left-side is greater than that of the right-side by one if the total number of amino acids in the loop is even, and both sides have the same number of amino acids when the total number of amino acids in the loop is odd.

Because the initial and terminal positions of the loops are strongly conserved (Yan and Sun, 1999), we also employed two other selection methods. The first method was that the end of the left loop must be located in the tenth position, and the other was that the beginning of the right loop was located in the twenty-third position.

If the length of the pattern was less than 32, the residues flanking the peptide in the primary amino acid sequence were appended at both ends (Kuhn et al., 2004). Examples of five fixed-length patterns are shown in Fig. 2.

We analyzed the conservation of 32 amino acid positions in the above-mentioned methods using the WEBLOGO server (http://weblogo.berkeley.edu/logo.cgi). Due to space constraints, only an example analysis of the aligned twenty-eighth position in the dataset is shown in Fig. 3.

The analysis indicated that both the βαβ motifs and non-βαβ motifs have strong amino acid conservation at various positions, and the types of motifs have different amino acid conservation in the same position. For example, at position 25, the most conserved amino acid is A in βαβ motifs, but the most conserved amino acid is G in non-βαβ motifs. Similarly at position 28, the most conserved amino acid is P in βαβ motifs, but the most conserved amino acid is D in non-βαβ motifs. At position 26, the most conserved amino acid is G in βαβ motifs, followed by L, K and P, but the most conserved amino acid is G, followed by K, E and D. G is more conserved in βαβ motifs than in non-βαβ motifs at position 27.

2.2. The selection of the features

2.2.1. Hydropathy compositions (Q)

Because the conformation and stabilization of proteins are seriously influenced by the hydropathy properties of their constituent amino acids (Hu et al., 2010), hydropathy composition was examined as a feature. The classification of hydropathy is

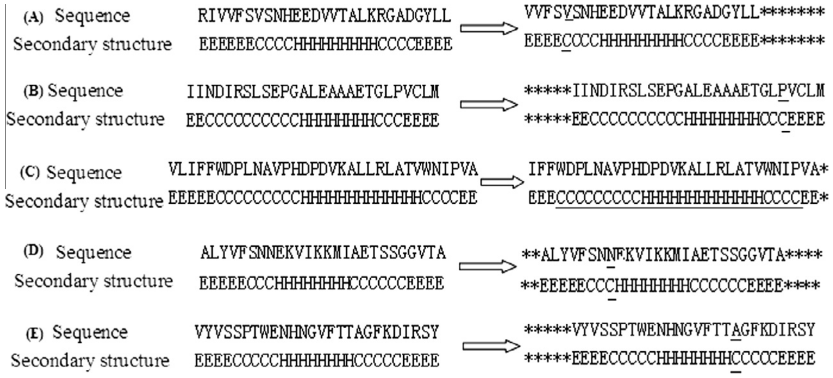
(A) Sequence
Secondary structure
RIVVFSVSNHEEDVVTALKRGADGYLL
EEEEEECCCCHHHHHHHHHCCCCEEEE
VVFSVSNHEEDVVTALKRGADGYLL*******
EEEECCCCHHHHHHHHHCCCCEEEE*******
(B) Sequence
Secondary structure
IINDIRSLSEPGALEAAAETGLPVCLM
EECCCCCCCCCCHHHHHHHHCCCEEEE
*****IINDIRSLSEPGALEAAAETGLPVCLM
*****EECCCCCCCCCCHHHHHHHHCCCEEEE
(C) Sequence
Secondary structure
VLIFFWDPLNAVPHDPDVKALLRLATVWNIPVA
EEEEECCCCCCCCCHHHHHHHHHHHHHCCCCEE
IFFWDPLNAVPHDPDVKALLRLATVWNIPVA*
EEECCCCCCCCCHHHHHHHHHHHHHCCCCEE*
(D) Sequence
Secondary structure
ALYVFSNNEKVIKKMIAETSSGGVTA
EEEEECCCHHHHHHHHCCCCCCEEEE
ALYVFSNNEKVIKKMIAETSSGGVTA**
EEEEECCCHHHHHHHHCCCCCCEEEE**
(E) Sequence
Secondary structure
VYVSSPTWENHNGVFTTAGFKDIRSY
EEEECCCCCHHHHHHHHCCCCCEEEE
*****VYVSSPTWENHNGVFTTAGFKDIRSY
*****EEEECCCCCHHHHHHHHCCCCCEEEE

Figure 2 Five examples of fixed-length patterns. Note: the sites which are emphasized are underlined; the symbol "*" shows residues flanking the peptide which were appended at both ends.

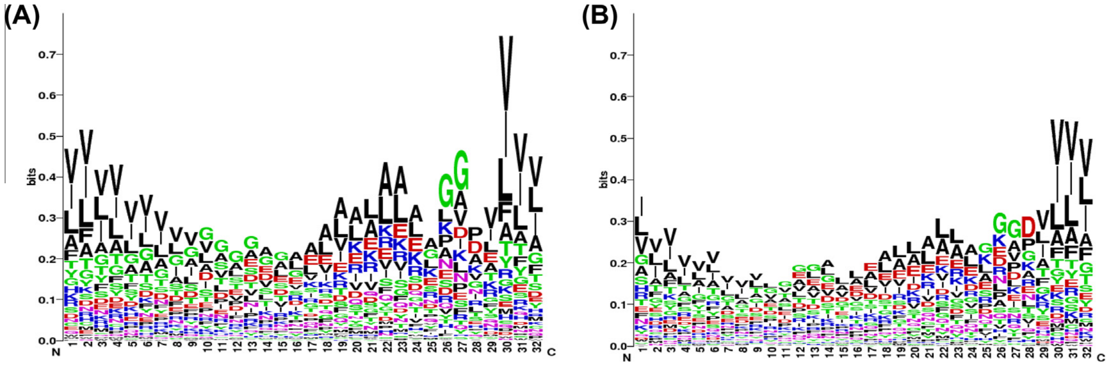

Figure 3 Sample of position conservation in the aligned twenty-eighth position. Note: βαβ motifs were shown in subfigure (A) and non-βαβ motifs were showed in subfigure (B). The overall height of the stack indicates position conservation, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position.

obtained from Pánek's scale (Pánek et al., 2005). The hydropathy composition was represented by a vector of 6 dimensions that contained the frequency of every class of amino acids.

2.2.2. Optimized amino acid composition of position (A)

Fig. 3 shows that βαβ motifs and non-βαβ motifs have strong position conservation. Thus, the position-specific amino acid composition was extracted as a feature of the fixed-length pattern for the βαβ motifs and non-βαβ motifs. This kind of feature was represented as a vector of 21 × 32 dimensions for every fixed-length pattern of the βαβ motifs and non-βαβ motifs (the 21-length dimension denotes 20 native amino acids and one terminal). This could produce vectors with a dimension of 3360 (21 × 32 × 5). The SVM algorithm may result in over-fitting phenomenon when high-dimensional features are used, and thus, we optimized the dimensions of the amino acid component of the position to avoid over-fitting phenomenon and improve the classification accuracy. In this paper, the mRMR (Maximum Relevance Minimum Redundancy) program (Li et al., 2012a) was used for optimizing the dimensions of the features.

The mRMR is a criterion of feature optimization proposed by Peng et al. (2005). The core idea of mRMR is to calculate the relevance between features and classes, and the redundancy between different features by using mutual information. Because of the excellent performance, mRMR has been widely used in many subjects including protein study field (Li et al., 2012a,b). In this work, we used the mRMR program as implemented in a software package. The mRMR program was used to optimize 3360 dimension features, and the top 100 features were selected, with a cumulative contribution rate up to 90%.

2.2.3. Predicted function motif (M)

Some conserved regions of the protein sequence are normally associated with biological function (Saha and Raghava, 2006), therefore, the predicted function motif was used as a feature.

PROSITE (http://www.expasy.org/tools/scanprosite/) was used in searching for the predicted function motifs from protein sequence data. Overall, 149 classes of predicted function motifs were obtained. We classified the predicted function motif into 3 categories. If the frequency was greater than 100, the motif was designated as a common function motif. If the frequency was greater than 10 and less than 100, the motif was designated as a general function motif. If the frequency was less than 10, the motif was designated as a scarce function motif.

Features of the predicted function motifs were represented as a three-dimensional vector in which each element represents the occurrence times about one of the function motif categories. If the motif was presented 1 time, then it was assigned a value of 1. If the motif presented 2 times, then it was assigned a value of 2. If it presented 3 times, then it was assigned a value of 3 and so on. Otherwise, it was assigned a value of 0.

2.2.4. Predicted secondary structure information (P)

Previous studies of predicting β-hairpin motifs (Kumar et al., 2005; Cruz et al., 2002) indicate that the predicted secondary structure information is an effective feature. Here we also extracted predicted secondary structure information for the sequence fragments of βαβ motifs and non-βαβ motifs. These data were obtained by using PSIPRED (Jones, 1999). The predicted secondary structure was represented by three-dimensional vectors with the elements representing the frequency of α-helix, β-sheet and the coil, respectively.

2.3. Methods

2.3.1. Support Vector Machine (SVM) algorithm

The basic idea of the SVM algorithm is to transform the input vector into a high-dimension Hilbert space and to seek an optimal hyperplane in this space, which maximizes the distance among the various samples. Moreover, it maximizes generalizability (Vapnik, 1995, 1998; Scholkopf et al., 1999).

The SVM algorithm is a convex optimization problem, and thus, a local optimal solution is the global optimal solution. The form of the decision function is as follows:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{k} a_i^* y_i k(x, x_i) + b^*\right) \quad (1)$$

where, $x \in R^n$, $y_i \in \{+1, -1\}$, a^*, is the Lagrange factor, and b^* is the classification thresholds, $k(x, x_i)$ is the kernel functions. Many successful predictions indicate that SVM is an effective method to predict the small sample. In addition, predicting protein structure by combining mRMR into SVM is a successful classification method (Liu et al., 2012).

The SVM algorithm has been compiled into software packages (Chang and Lin, 2011). In this work, we used the libsvm-3.0 package and selected the radial kernel functions (RBF) $k(x, x_i) = \exp\left(-g\|x - x_i\|^2\right)$. The above features are inputted into SVM for training. We obtained optimal C and gamma values of 32 and 0.089285, respectively. Lastly, a classifier was constructed, and was used to predict βαβ motifs in the test dataset.

2.3.2. Performance measure

We used the following standard measures (Hu et al., 2010; Hu and Li, 2008) to evaluate the performance: sensitivity of βαβ motifs ($Sn_{\beta\alpha\beta}$), sensitivity of non-βαβ motifs ($Sn_{n\beta\alpha\beta}$), specificity of βαβ motifs ($Sp_{\beta\alpha\beta}$), specificity of non-βαβ motifs ($Sp_{n\beta\alpha\beta}$), Matthew's correlation coefficient (*MCC*), and accuracy of prediction (*Acc*), which was calculated with the following formula:

$$Sn_{(\beta\alpha\beta)} = \left[\frac{TP}{(TP + FN)}\right] \times 100\% \tag{2}$$

$$Sn_{(\mathrm{n}\beta\alpha\beta)} = \left[\frac{TN}{(TN + FP)}\right] \times 100\% \tag{3}$$

$$Sn_{(\beta\alpha\beta)} = \left[\frac{TP}{(TP + FP)}\right] \times 100\% \tag{4}$$

$$Sn_{(\mathrm{n}\beta\alpha\beta)} = \left[\frac{TN}{(TN + FN)}\right] \times 100\% \tag{5}$$

$$MCC = \frac{[(TP \times TN) - (FP \times FN)]}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

$$Acc = \left[\frac{(TP + TN)}{(TP + FP + TN + FN)}\right] \times 100\% \tag{7}$$

Here, *TP* and *TN* denote the number of correctly predicted βαβ motifs and non-βαβ motifs, respectively. *FN* denotes the number of the βαβ motifs that are predicted as non-βαβ motifs, and *FP* denotes the number of the non-βαβ motifs that are predicted as βαβ motifs.

2.3.3. Test method

In this work, we also employed 5-fold cross-validation to evaluate the performance of our method. The dataset was randomly divided into five sets, and each set contained an equal number of βαβ motifs, with four of them used for the training set and the fifth used as a test dataset. The process was repeated five times with a different set for validations each time. The final predictive results were averaged over the five test sets.

Moreover, independent test can reflect the robustness of the classifier. Therefore, in this paper, we also used the independent test, in which the proteins in the test set were not presented in the training set.

3. Results and discussion

We selected study objects where the loop-α-loop length was from 10 to 26 amino acids. The ideal fixed-length pattern comprised 32 amino acids. The SVM algorithm was applied to predict βαβ motifs.

Table 2 Predictive results using a 5-fold cross-validations test.

	$Sn_{(\beta\alpha\beta)}$ (%)	$Sn_{(\mathrm{n}\beta\alpha\beta)}$ (%)	$Sp_{(\beta\alpha\beta)}$ (%)	$Sp_{(\mathrm{n}\beta\alpha\beta)}$ (%)	*MCC*	*Acc* (%)
Q	74.1	52.9	66.9	61.5	0.28	64.8
Q + *A*	80.6	75.0	80.5	75.1	0.55	78.1
Q + *A* + *M*	83.5	74.8	81.0	78.0	0.59	79.7
Q + *A* + *M* + *P*	85.4	77.0	82.6	80.4	0.63	81.7

3.1. The predictive results using 5-fold cross-validation test

We applied the method to predict βαβ motifs by adding the features successively. The predictive results using 5-fold cross-validation are shown in Table 2.

When the hydropathy composition (*Q*) was used as a feature, the overall accuracy of the prediction was only 64.8%. When the optimized amino acid composition of position (*A*) was added, *Acc* and *MCC* values of 78.1% and 0.55, respectively, could be achieved. The sensitivity and specificity was also greatly increased, with values greater than 75.0%. To improve the prediction accuracy, we added the predicted function motif (*M*) and predicted secondary structure information (*P*). The results were further improved when both features are used. We observed that the best predictive result was obtained by using the composite vector *Q* + *A* + *M* + *P*, wherein *Acc* and *MCC* values of 81.7% and 0.63, respectively, were achieved, and the value of sensitivity and specificity was more than 77.0%. Apparently, all information concerning βαβ motifs cannot be expressed only by using sequence information. The predictive results were greatly improved if the predicted function motif and predicted secondary structure information were both used as features.

3.2. Predictive results using independent test

The purpose of protein structure prediction is to predict the unknown structure protein chain. Therefore, in this work, we also predicted the βαβ motif by using the independent test. The 1200 protein chains were randomly selected as the training dataset, which contained 3080 βαβ motifs and 2362 non-βαβ motifs. The remaining 240 protein chains were used as the independent test dataset, which contained 604 βαβ motifs and 511 non-βαβ motifs.

Prediction results of the independent test using the composite vector *Q* + *A* + *M* + *P* are shown in Table 3. The *Acc* and *MCC* values were 76.7% and 0.53, respectively. It can be clearly seen that the predictive results of the 5 cross-validation test and independent test are consistent.

Fig. 4 an example Prediction of βαβ motif (from the protein chain 1EDZ). This protein contains a total of 4 ECHCE patterns: two βαβ motifs (colored with red and yellow) and

Table 3 Predictive results using independent tests.

$Sn_{(\beta\alpha\beta)}$	$Sn_{(\mathrm{n}\beta\alpha\beta)}$	$Sp_{(\beta\alpha\beta)}$	$Sp_{(\mathrm{n}\beta\alpha\beta)}$	*MCC*	*Acc*
81.6%	70.8%	76.8%	76.5%	0.53	76.7%

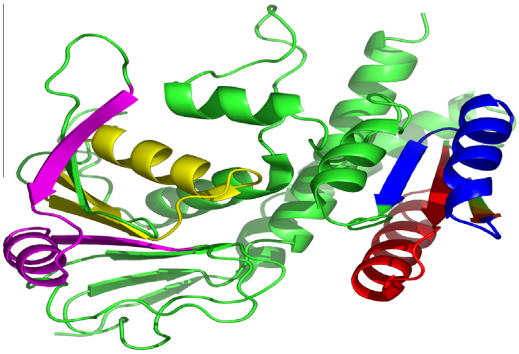

Figure 4 Prediction βαβ motif (from the protein chain 1EDZ) example.

two non-βαβ motifs (colored with blue and magenta). The first three motifs have been correctly predicted, and the last one has been wrongly predicted as the βαβ motif.

4. Conclusions

In this paper, we constructed a dataset for a complex super-secondary structure βαβ motif. The dataset contained 4277 βαβ motifs and 3366 non-βαβ motifs. Based on statistical analysis and biology background of the motif, we selected hydropathy composition, optimized amino acid composition of position, predicted function motif and predicted secondary structure information as features, they reflected comprehensive sequence characteristics. In this study we have found that predicting results continually improved with adding features, the best predictive result was obtained by using the combined vector $Q + A + M + P$, no matter using the 5 cross-validation test or independent test. Therefore the SVM algorithm based on combined vector is an effective method to predict βαβ motifs. In further work, we will consider to add other helpful features such as hydrogen bonds to obtain an improved predict performance. In the algorithm aspect we will try to use the Random Forest algorithm, and look forward to improve the predicted results.

Conflict of interest

No conflict of interest exits in the submission of this work, and work is approved by all authors for publication.

Acknowledgments

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

References

Blundell, T.L., Sibanda, B.L., Sternberg, M.J., Thornton, J.M., 1987. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326, 347–352.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27.

Chou, K.C., 1995. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins Struct. Funct. Genet. 21, 319–344.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Cruz, X., Hutchinson, E.G., Shepherd, A., Thornton, J.M., 2002. Toward predicting protein topology: an approach to identifying β hairpins. Proc. Natl. Acad. Sci. 99, 11157–11162.

Hu, X.Z., Li, Q.Z., 2008. Prediction of the β-hairpins in proteins using support vector machine. Protein J. 27, 115–122.

Hu, X.Z., Li, Q.Z., Wang, C.L., 2010. Recognition of β-hairpin motifs in proteins by using the composite vector. Amino Acids 38, 915–921.

Jia, S.C., Hu, X.Z., 2011. Using random forest algorithm to predict β-hairpin motifs. Protein Pept. Lett. 18, 609–617.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

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

Kuhn, M., Meiler, J., Baker, D., 2004. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins Struct. Funct. Bioinf. 54, 282–288.

Kumar, M., Bhasin, M., Natt, N.K., Raghava, G.P.S., 2005. BhairPred: prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucl. Acids Res. 33, 154–159.

Li, B.Q., Huang, T., Liu, L., Cai, Y.D., Chou, K.C., 2012a. Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network. PLoS One 7, e33393.

Li, B.Q., Hu, L.L., Niu, S., Cai, Y.D., Chou, K.C., 2012b. Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. J. Proteomics 75, 1654–1665.

Liu, L., Hu, X.Z., Liu, X.X., Wang, Y., Li, S.B., 2012. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions. Protein Pept. Lett. 19, 439–449.

Pánek, J., Eidhammer, I., Aasland, R., 2005. A new method for identification of protein (Sub) families in a set of proteins based on hydropathy distribution in proteins. Proteins Struct. Funct. Bioinf. 58, 923–934.

Peng, H.C., Long, F.H., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238.

Saha, S., Raghava, G.P.S., 2006. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucl. Acids Res. 34, 202–209.

Scholkopf, B., Burges, C.J.C., Smola, A.J., 1999. Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA.

Sun, Z.R., Rao, X.Q., Peng, L.W., Xu, D., 1997. Prediction of protein supersecondary structures based on the artificial neural network method. Protein Eng. 10, 763–769.

Taylor, W.R., Thornton, J.M., 1983. Prediction of super-secondary structure in proteins. Nature 301, 540–542.

Taylor, W.R., Thornton, J.M., 1984. Recognition of super-secondary structure in proteins. J. Mol. Biol. 173, 487–512.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Wierenga, R.K., Terpstra, P., Hol, W.G.J., 1986. Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint. J. Mol. Biol. 187, 101–107.

Yan, L.F., Sun, Z.R., 1999. Molecule Structure of Protein. Tsinghua University Press.