Saudi Journal of Biological Sciences (2017) 34, 1361–1369

King Saud University

Saudi Journal of Biological Sciences

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

ORIGINAL ARTICLE

Using feature optimization-based support vector machine method to recognize the β-hairpin motifs in enzymes

Dongmei Li, Xiuzhen Hu *, Xingxing Liu, Zhenxing Feng, Changjiang Ding

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

Received 5 September 2016; revised 16 November 2016; accepted 17 November 2016
Available online 28 November 2016

KEYWORDS

Enzymes;
β-Hairpin motif;
Ligand binding site;
Support vector machine;
Minimum redundancy maximum

Abstract β-Hairpins in enzyme, a kind of special protein with catalytic functions, contain many binding sites which are essential for the functions of enzyme. With the increasing number of observed enzyme protein sequences, it is of especial importance to use bioinformatics techniques to quickly and accurately identify the β-hairpin in enzyme protein for further advanced annotation of structure and function of enzyme. In this work, the proposed method was trained and tested on a non-redundant enzyme β-hairpin database containing 2818 β-hairpins and 1098 non-β-hairpins. With 5-fold cross-validation on the training dataset, the overall accuracy of 90.08% and Matthew's correlation coefficient (Mcc) of 0.74 were obtained, while on the independent test dataset, the overall accuracy of 88.93% and Mcc of 0.76 were achieved. Furthermore, the method was validated on 845 β-hairpins with ligand binding sites. With 5-fold cross-validation on the training dataset and independent test on the test dataset, the overall accuracies were 85.82% (Mcc of 0.71) and 84.78% (Mcc of 0.70), respectively. With an integration of mRMR feature selection and SVM algorithm, a reasonable high accuracy was achieved, indicating the method to be an effective tool for the further studies of β-hairpins in enzymes structure. Additionally, as a novelty for function prediction of enzymes, β-hairpins with ligand binding sites were predicted. Based on this work, a web server was constructed to predict β-hairpin motifs in enzymes (http://202.207.29.251:8080/).

1. Introduction

Super secondary structure is a building block of the tertiary structure of protein, and this geometrical arrangement of the local space structure was constructed by two or several secondary structure units that are connected by loop. In definition of β-hairpin patterns, an adjacent anti-parallel β-strand connects with another by one or more hydrogen bonds; otherwise, the patterns were called non-β-hairpins (Kuhn et al., 2004).

* Corresponding author.
E-mail address: hxz@imut.edu.cn (X. Hu).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.sjbs.2016.11.014

Because β-hairpin is a simple arrangement of the β-strand and includes rich folding information, correctly identifying β-hairpin will contribute to fold recognition and structure assembly (Jenny et al., 1995; Wintjens et al., 1996). In recent decades, varied studies of theoretical prediction on β-hairpin have been developed. In 2002, the artificial neural network (ANN) was employed to predict β-hairpins contained in 534 proteins with a prediction accuracy of 47.7% (Cruz et al., 2002).

In 2004, an ANN algorithm was applied to identify local hairpins and non-local diverging turns from 2209 proteins, and an accuracy of 75.9% was obtained (Kuhn et al., 2004). Then the support vector machine (SVM) was used to predict β-hairpins in a database of 2880 proteins (EVA), and an accuracy of 79.2% (with a Mcc of 0.59) was achieved (Kumar et al., 2005).

In 2008, based on composite vector, SVM was applied to predict β-hairpins in ArchDB40 (including 3088 proteins) and EVA database, the accuracies of cross-validation and independent testing were 79.9% and 83.3%, and the corresponding Mcc values were 0.59 and 0.67, respectively (Hu and Li, 2008a). In 2010, a method of quadratic discriminant (QD) with improved composite vector was developed to predict β-hairpins in ArchDB40 and EVA database (Hu et al., 2010). With a 5-fold cross-validation and independent test, the overall accuracies reached to 83.1% (with the Mcc values of 0.59) and 80.7% (with the Mcc values of 0.61), respectively. In 2013, Random Forest algorithm was applied to predicted β-hairpin motifs in ArchDB40 dataset, based on 5-fold cross-validation, and the overall accuracy was up to 83.3% (with Matthew's correlation coefficient of 0.59). Additionally, with the same features and testing method, SVM algorithm was used as a comparison with the Random Forest; however, the prediction performance was not so well. (Jia et al., 2013). In 2015, based on the chemical shifts, an algorithm called quadratic discriminant was developed to identify beta-hairpin motifs, and the prediction results with sensitivity of 92%, the specificity of 94%, and Mathew's correlation coefficient of 0.85 were obtained (Feng and Kou, 2015).

Previous studies of β-hairpins prediction were based on all kinds of proteins. However, β-hairpins in different kinds of proteins have their particular properties, especially in enzymes protein. There is no doubt that the processes of digestion, absorption, respiration, motion and reproduction in organism all belong to enzymatic reaction. Almost all of the chemical reactions of metabolism in cell are catalysis of enzymes. Meanwhile, enzymes are also the critical important structure with known drug targets. All functions of enzymes, including signals relay, transport and catalysis, rely on the other molecules combined with enzymes, namely ligands. With binding ligands, enzyme can perform and regulate its functions directly, stabilize structure and lead to changes of conformation in order to influence the microenvironment, and in turn to control the protein functions indirectly. In enzymatic reactions, the ligand conformation ally fit into the ligand binding sites of the enzymes, which plays a critical role in controlling the spatial arrangements and orientations of the substrates in the active site. And the ligand specificities of enzymes are determined by these conformational restrictions. β-hairpin is simple arrangement of the β-strand, and a cooperative interaction between the two strands of the β-hairpin loop often plays important role in ligand binding of enzyme, for example, divergent β-hairpins in proximity of the active sites of ABH2 and ABH3 are central for substrate specificities. Swapping hairpins between the enzymes resulted in hybrid proteins resembling the donor proteins (Lee et al., 2005). For another example, remarkable binding ligands including FAD, ATP, NAD and metal ions Zn^{2+}, Ca^{2+}, Mg^{2+}, etc. are also contained in β-hairpin of enzyme proteins. FAD, the coenzyme of oxidoreductase, is involved in several important metabolic reactions of carbohydrate and lipid and amino acid. In tricarboxylic acid cycle, when accepting protons and turning into FADH2, FAD is oxidized as FAD+ in the respiratory chain (Stryer et al., 2011). NAD is the coenzyme of dehydrogenase. When acting on the CH-OH group of donor with NAD+ or $NADP^+$ as acceptor, it will result in the enzymic reaction of glycerophospholipid metabolism (Edgar and Bell, 1978). Zn^{2+} acts as the role of Lewis acid in pancreatic carboxypeptidase which belongs to lyase, and the inductive effect of attracting electrons makes the local substrate present positive electricity. Thus, it is easy for OH^- or H_2O to nucleophilic attack with substrate, and lead to the hydrolysis of substrate. So Zn^{2+} is important for the biological process of protein hydrolysis (Fruton, 1999). Because enzymes have their own properties and the β-hairpins in enzymes often contain ligand binding sites, the prediction of β-hairpins in enzyme protein would be more significant. In this paper, an effort was made to achieve this purpose.

A total of 2818 β-hairpins and 1098 non-β-hairpins in enzymes protein were obtained as research objects. Six groups of features were extracted from the information of original sequence and predicted secondary structure. After the optimization of the original features by the criterion of minimum redundancy maximum relevance (mRMR), 245 out of 906 original features were selected and input into SVM for prediction. Experimental results show that the selected features can achieve the best performance. Additionally, our method was used to predict the 845 β-hairpins containing ligand binding sites, and good results were obtained.

2. Materials and methods

2.1. Materials

2.1.1. Enzyme β-hairpin database

As the classification of the structure of protein loops, ArchDB database (http://sbi.imim.es/cgi-bin/archdb/loops.pl) was generated from proteins with known structure. The data were derived from DSSP (Sander and Kabasch, 1983) and reorganized by Oliva et al. (1997), Espadaler et al. (2004) and Bonet et al. (2014). According to the regulation secondary structures connected by loops, the super secondary structures can be classified into five types: alpha-alpha, beta-beta link, beta-beta hairpin, alpha-beta and beta-alpha. Among them, beta-beta hairpin was taken as beta-hairpin and beta-beta link as non-β-hairpin (Hu and Li, 2008a; Hu et al., 2010). ArchDB database contained four sub-datasets: ArchDB_95, ArchDB_40, ArchDB_EC and ArchDB_KI, which has been previously used to predict β-hairpins. In this work, ArchDB_EC was selected, which contains protein chains with known enzyme function and the structure resolution < 3.0 Å, among which arbitrary two sequences have a percentage identity about 75%. The non-redundant Enzyme β-hairpin database was constructed as the following steps:

I. 1781 protein chains 'PDB-ID' were obtained from ArchDB_EC, among which each had more than one β-hairpin. II. The structures of the 1781 protein chains were extracted from PDB (http://www.rcsb.org/pdb/). III. By using BLAST software (Tatusova and Madden, 1999) to filter the redundant sequences from the 1781 protein, 1080 protein chains were reserved, and the sequence identity between each two proteins was not higher than 25%. According to international enzyme classification, the 1080 protein chains belong to 7 types, and the number of proteins in each type was as follows: 1. Oxidoreductase (200), 2. Transferase (266), 3. Hydrolase (331), 4. lyase (76), 5. Isomerase (49), 6. Ligase (55), 7. The others (103) (mutase, tyrosine kinase, etc.). (http://202.207.29.251:8080/) IV. 2846 β-hairpins and 1186 non-β-hairpins were obtained from the 1080 protein sequences. Among these β-hairpins, 861 motifs contained ligand binding sites.

A statistical analysis was made on the 2846 β-hairpins and 1186 non-β-hairpins. As shown in Fig. 1, the shortest and longest loop lengths for β-hairpins and non-β-hairpins were 1 and 32, respectively. About 97% of the original motifs have the patterns with loop length of 2–12, and then this portion was reserved as the research object. Overall, 2818 β-hairpins and 1098 non-β-hairpins were reserved, accounting for 99% and 92% of the original motifs, respectively. Within the reserved 2818 β-hairpins, 845 β-hairpins contain ligand binding sites, which account for 98% of the original 861β-hairpins with ligand binding sites.

Note: The abscissa represents different loop lengths and ordinate represents the number of motifs with different loop lengths. The dark and gray histograms represent the distributions of β-hairpins and non-β-hairpins, respectively.

2.1.2. Experimental enzyme β-hairpin database

To test the prediction ability of our approach, a dataset independent from ArchDB_EC database was built and the processes were as follows.

I. 89 proteins' PDB-ID containing 306 chains with structure resolution < 3.0 Å were randomly selected from ENZYME (http://enzyme.expasy.org/). II. The structures of the 306 protein chains were extracted from PDB database.III. BLAST software was used to filter the redundant proteins, and 110 protein chains were kept at last, in which the sequence identity of arbitrarily protein chain with another was below 25%. IV. DSSP software was used to assign secondary structure to each amino acid (Sander and Kabasch, 1983), where the DSSP labels of 'H', 'G' and 'I' were converted as α-helix(H), 'E' and 'B' as β-strand(E), 'T', 'S' and ' '(space) as coil(C). 525 ECE (β-strand coil β-strand) patterns were obtained by secondary structure assignment from DSSP. The number of patterns with loop length of 2–12 was 448. V. PROMOTIF software (Hutchinson and Thornton, 1996) was used to locate β-hairpins in the 110 protein chains. Among the 448 patterns, 228 were assigned as β-hairpins by PROMOTIF; the rest 220 patterns were assigned as non-β-hairpins.

2.2. Methods

2.2.1. Feature extraction

The average pattern length of β-hairpins and non-β-hairpins was 14.9 and 13.4, respectively. Following the guideline of previous studies (Hu and Li, 2008a), the pattern length with 15 amino acids residues was selected as the best fixed-length pattern. For each β-hairpin and non-β-hairpin, the fixed-length pattern was generated using the scheme described below: Set loop as the center of the pattern; If length of pattern was less than 15, we appended the residues flanking the peptide in the primary sequence at both ends; If the value of loop length was even, the loops of left-hand side keep one more amino acid residue than those of right-hand side.

Referring to our group's studies (Hu and Li, 2008a; Hu et al., 2010), amino acid composition was an efficient parameter for identifying β-hairpins. Also, amino acid dipeptide composition was also powerful feature for it can represent the correlation between two adjacent amino acids. Moreover, predicted secondary structure and hydropathy characteristic classification for amino acids have been commonly utilized in the identification of β-hairpins as parameters. These parameters were beneficial to promote the prediction results. In order to collect as much classify information as possible, six groups of features to represent identification information were extracted by two the following methods.

2.2.2. Original feature extraction based on the best fixed-length patterns

Three groups of parameters were extracted here: amino acid compositions of each position ($21 * 15 = 315$, 21 include 20 types amino acid and one terminal residues), hydropathy

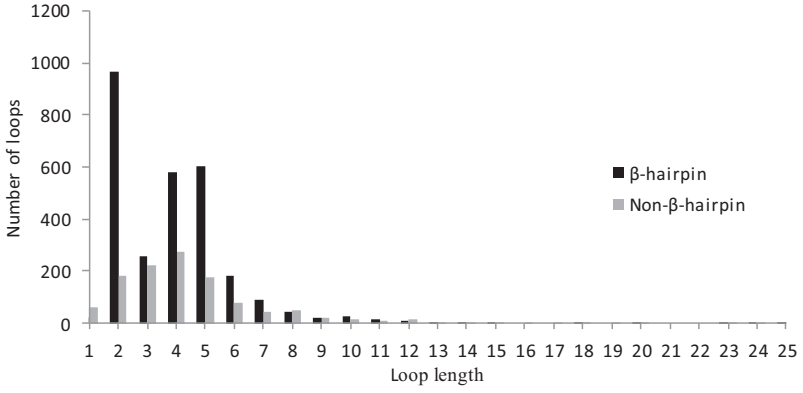

Fig. 1 The distribution of the numbers of motifs with different loop lengths.

characteristics for amino acid of each position (7 * 15 = 105) and predicted secondary structures of each position (4 * 15 = 60).

2.2.3. Original feature extraction based on the original patterns

Within this approach, another three groups of features were extracted: amino acid composition (20), hydropathy characteristics for amino acid (6) and amino acid contiguous dipeptides composition (400).

Taken together, a total of 906 features were extracted for prediction. Three features of predicted secondary structure were from PSIPRED (McGuffin et al., 2000) (http://bioinf.cs.ucl.ac.uk/psipred/), which predict secondary structure information from original sequences. PSIPRED outputs E, H and C represent β-strand, α-helix and coil, respectively. The 6 features of hydropathy characteristics (Pánek et al., 2005) are described in Fig. 2.

2.2.4. Feature optimization

Feature optimization is a key issue in pattern classification, which significantly influences the prediction power of one classifier. Protein sequence information can be represented by multidimensional features, but there were many redundant or irrelevant features, which may make it difficult to construct a classifier. Hence, to improve the prediction performance, the primary goals of feature optimization were to optimize predictive characters, remove noise, reduce feature dimension and avoid over fitting.

mRMR (Maximum Relevance Minimum Redundancy) algorithm is a criterion of features optimization proposed by Peng et al. (2005). The core idea of mRMR is to calculate the relevance between features and classified targets and the redundancy between different features by using mutual information.

Suppose there are two random variables X and Y. Their probability densities are $P(x)$ and $P(y)$ and joint probability density is $P(x, y)$. The mutual information value between X and Y is calculated using the following equation:

$$I(x,y) = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (1)$$

According to the maximum relevance criterion, the mutual information value of feature x_i with the target class C should be maximum. The top m features that have the maximum mutual information values with target classes usually are selected as feature subset. The maximum relevance is defined as follows:

$$\max(D),\ D = \frac{1}{|S|} \sum_{x_i \in S} I(X_i, C) \quad (2)$$

where D represents the relevance of the subset S with m features.

However, there are still many redundant features in the subset selected by maximum relevance criterion. When a feature highly depends on another and one was removed, the class-discriminative power would not change obviously. Therefore, it is necessary to take the minimum redundancy criterion based on the maximum relevance of features into consider. The minimum redundancy is defined as follows:

$$\min(R),\ R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(X_i, X_j) \quad (3)$$

Combining the above two criteria, mRMR optimization criterion has the following simple form:

$$\max(\Phi),\ \Phi = D - R \quad (4)$$

In our study, we used the criterion of mRMR to filter the 906 features extracted from β-hairpins and non-β-hairpins. The value of Φ for each feature was obtained and sorted. Depended on the abundant prediction results, the prediction gets the best performance when reserving the top 245 features. Table 1 shows the selected features as follows.

Table 1 The number of features of six groups after selection by mRMR.

Feature	Original number	Selected number
1. AACP	315	74
2. HCP	105	30
3. PSSP	60	23
4. ACC	20	5
5. HC	6	4
6. AACD	400	109
Total	906	245

AACP: amino acid compositions of each position; HCP: hydropathy characteristics for amino acid of each position; PSSP: predicted secondary structures of each position; ACC: amino acid composition; HC: hydropathy characteristics for amino acid; AACD: amino acid contiguous dipeptides composition.

2.2.5. Support vector machine

As a machine learning algorithm proposed by Vapnik (1995, 1998), SVM has been proposed in many previous reports, such as protein structure prediction (Hu and Li, 2008a, 2008b), protein sub-cellular localization (Chou and Cai, 2002) and classification of protein folding (Ding and Dubchak, 2001; Shi et al., 2006; Liu et al., 2012). SVM algorithm searches for a linear separating hyperplane with the maximal margin, and ensures accuracy of classification as well. The minimal error classification model generated by SVM through training dataset of definite samples can guarantee the same performance for independent testing dataset. To extend SVM from linear filed to nonlinear, Vapnik (1995, 1998) map input features into a higher dimensional Hilbert space by using kernel function

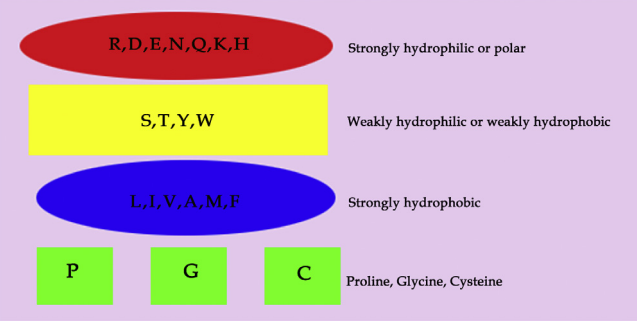

Fig. 2 Hydropathy characteristics for amino acids.

Table 2 The predictive results with different dimensions of features selected by mRMR.

Dimension	*Acc* (%)	*Mcc*	S_nH (%)	S_nNH (%)	S_pH (%)	S_pNH (%)
20	86.97	0.67	92.49	72.81	89.72	79.08
50	86.59	0.66	91.43	74.18	90.08	77.13
100	87.28	0.68	92.17	74.72	90.34	78.81
150	89.04	0.72	93.98	76.36	91.07	83.18
200	89.96	0.74	94.94	77.18	91.44	85.60
245	90.08	0.74	95.47	76.22	91.15	86.78
300	89.77	0.73	95.47	75.13	90.78	86.61
350	89.73	0.73	96.22	73.08	90.17	88.28
400	89.19	0.74	97.01	72.50	88.97	91.41
450	88.16	0.69	97.87	63.25	87.23	92.04
500	84.45	0.59	98.88	47.40	82.83	94.29

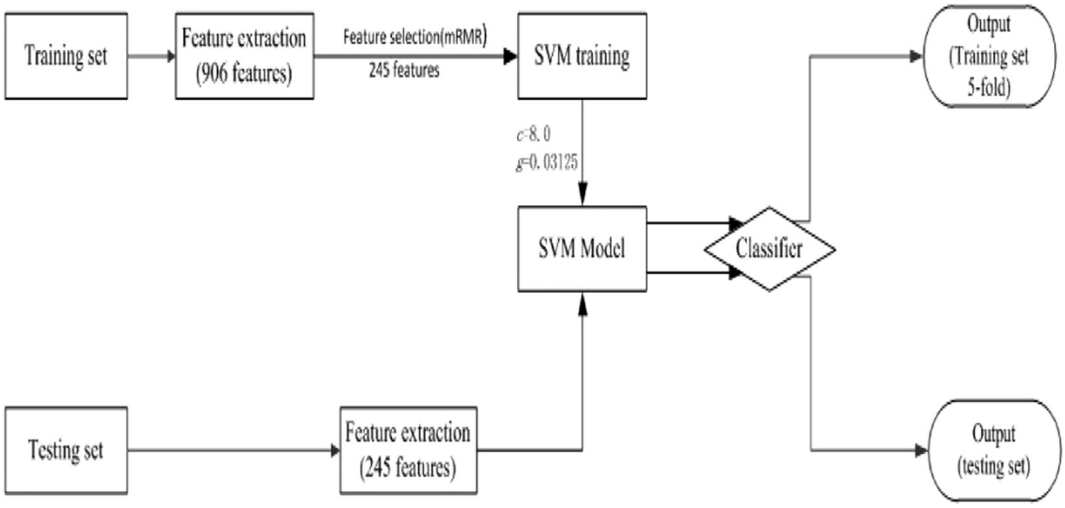

Fig. 3 Flowchart of the prediction process for 5-fold cross-validation and independent test.

and then construct optimal hyperplane in this space. The calculating formulation of optimal hyperplane is shown below:

$$h(x) = \mathrm{sgn}\left(\sum_{j=1}^{k} a_j k(X, X_i) + b\right) \tag{5}$$

where $k(X, X_i)$ is called the kernel function. It generally has the following four types: Liner;

$$kX_iX_j = X_i^T X_j \tag{6}$$

Polynomial; Radial basis function (RBF); Sigmoid.

$$K(X_i, X_j) = \left(X_i^T X_j + 1\right)^d \tag{7}$$

$$K(X_i, X_j) = \exp\left(-g\left\|X_i - X_j\right\|^2\right) \tag{8}$$

$$K(X_i, X_j) = \tanh\left(X_i^T X_j\right) + c \tag{9}$$

SVM has been implemented as software by many researchers, such as libsvm, mysvm and svmlight. Here libsvm-2.93 package (http://www.Csie.ntu.edu.tw/cjlin/libsvm) was used and RBF was chosen as the kernel function in calculation. The top 245 features selected by mRMR were input into SVM after scaling the values of features in training dataset, and then an approach of gird-search was used to determine the best value of C (8.0) and gamma (0.03125) parameters. Finally a classifier was established. This classifier was used to predict β-hairpins and non-β-hairpins in the testing dataset and evaluate its ability of generalization.

2.2.6. Performance measures

This paper used standard measures adopted by previous studies of β-hairpins prediction to estimate the performance of our method: accuracy of prediction (*Acc*), Matthews' correlation coefficient (*Mcc*), sensitivity of β-hairpin (S_nH), sensitivity of non-β-hairpin (S_nNH), specificity of β-hairpin (S_pH), and specificity of non-β-hairpin (S_pNH). Above values were calculated by the following:

$$Acc = \frac{p + r}{p + r + o + u} \tag{10}$$

$$Mcc = \frac{p \times r - o \times u}{\sqrt{(p + o)(p + u)(r + o)(r + o)}} \tag{11}$$

$$S_nH = \frac{p}{p + u} \tag{12}$$

$$S_nNH = \frac{r}{r + u} \tag{13}$$

Table 3 The prediction results for 5-fold cross-validation and independent test.

	Acc (%)	*Mcc*	S_nH (%)	S_nNH (%)	S_pH (%)	S_pNH (%)
Training dataset	90.08	0.74	95.47	76.22	91.15	86.78
Testing dataset	88.93	0.76	90.61	85.18	93.16	80.29
Hu's (ArchDB)	83.1	0.59	91.3	64.3	85.4	76.4
Hu's (EVA)	80.7	0.61	83.4	77.4	81.8	79.3

Table 4 The testing results of β-hairpins in the enzyme experimental sequence dataset.

	Acc (%)	*Mcc*	S_nH (%)	S_nNH (%)	S_pH (%)	S_pNH (%)
DSSP	85.93	0.74	79.15	97.57	98.24	73.18
PSIPRED	70.67	0.41	73.07	68.64	66.27	75.14

$$S_pH = \frac{p}{p+o} \tag{14}$$

$$S_pNH = \frac{r}{r+u} \tag{15}$$

Here p and r denote the number of correctly predicted sequence segments for β-hairpins and non-β-hairpins, respectively. u donates the number of β-hairpins segments predicted as non-β-hairpins, o donate the number of non-β-hairpins predicted as β-hairpins.

3. Results and discussion

3.1. Prediction for β-hairpins in enzymes

2818 β-hairpins and 1098 non-β-hairpins were randomly divided into training dataset (1879 β-hairpins and 732 non-β-hairpins) and testing dataset (939 β-hairpins and 366 non-β-hairpins). mRMR criterion optimized 906 original features from information of sequence and predicted secondary structure.

The mRMR can obtain serial subsets comprising features sorted by the values of Φ. When selecting the subsets with top n features, the predictive results will be different. In this paper, denoting the number of features, the value of n was between 20 and 500. The top n features were inputted into SVM for prediction. Finally, the predicted results by using 5-fold cross-validation on training dataset were obtained. Some of prediction performance was shown as follows (Table 2).

It can be seen that the predicted results were optimum when the number of selected features was 245, and higher or lower number of features will result in declining performance, demonstrating the importance of feature optimization. So these 245 optimal features were used as the final predictive features.

The flowchart of the prediction process of 5-fold cross-validation for training dataset and independent test for testing dataset is shown in Fig. 3. Table 3 shows the prediction performance.

Note: 906 original features from training dataset were extracted, and then 245 features were selected by mRMR.

With 5-fold cross-validation, the optimum features were input into SVM. A classifier was established with a training model and through 5 times circulation, an output of 5-fold cross-validation for training set was obtained. Then 906 original features and optimized 245 features by mRMR from testing dataset were obtained in the same way. Based on the predictive model obtained from training set, 245 features from testing dataset were input into the SVM classifier for independent test. At last an output of testing set was obtained.

The predicted results show that on training dataset with 5-fold cross-validation, the accuracy was 90.08%, Mcc was 0.74, and the sensitivity and the specificity for β-hairpin were 95.47% and 91.15%, respectively. The prediction accuracy and Mcc of independent test on testing dataset were 88.93%, 0.76, respectively. The sensitivity and the specificity for β-hairpin were 90.61% and 93.16%, respectively.

As our method was developed to predict β-hairpins in enzymes for the first time, there was no comparison with previous studies. But we listed the best results of Hu et al. (2010) using QD method to predict β-hairpins without considering the kinds of proteins, with a 5-fold cross-validation on ArchDB_40 dataset, the accuracy was 83.1%, Mcc was 0.59, on EVA dataset, the accuracy was 80.7%, and Mcc was 0.61. It can be obviously seen that the performances obtained were better than those of Hu et al.

3.2. Prediction for β-hairpins on an enzyme experimental sequence dataset

In order to test the predictive ability of our method in real condition, the proposed method was tested on a dataset of β-hairpins and non-β-hairpins in an enzyme experimental sequences dataset built by our group. This dataset contains 228 β-hairpins and 220 non-β-hairpins assigned by DSSP and PROMOTIF software, which was used as independent testing dataset. The prediction model was constructed by using the former 2818 β-hairpins and 1098 non-β-hairpins as training dataset, and the model was then used to predict the β-hairpin from the experimental sequences. The accuracy was 85.93% with Mcc of 0.74, and the sensitivity and the specificity for β-hairpin were 79.15% and 98.24%, respectively (Table 4).

Actually, it is known that many enzyme proteins only have sequence information while with no observed secondary structure information, so we used predicted secondary structure to get the ECE pattern. In this way, 430 ECE patterns were

```
>1OID (chain: A)
1. IGNHEFDNPLTVLRQQEKWAKFPLLSANIYQKSTGERLFKPWALFKRQDLKIAVIGLTTDDTAKIGNPEYFTDIEFRKPA
2.   GGGGSS HHHHHHHHHH SS EE SSEEETTT  BSS  EEEEEETTEEEEEEEEE TTHHHHS GGGGTTEEE  HH
3. CCCCCCCCCCHHHHHHHHHHCCCCEEEEECCCCCCCCCCCCCCCEEEEEECCEEEEEEEECCCCCCCCCCCCCCCCCEEECHH
4.                                          WALFKRQDLKIAVIGL β#*
5.                         LLSANIYQKSTGERLFKPWALFKR* LKIAVIGLTTDDTAKIGNPEYFTDIEFR

1. DEAKLVIQELQQTEKPDIIIAATHMGHYDNGEHGCNAPGDVEMARALPAGSLAMIVGGHSQDPVCMAAENKKQVDYVPGT
2. HHHHHHHHHHHHTT  SEEEEEEES   GGG  TTS   HHHHHHHS TTSSSEEE  SS  B  EEETTEE SS  TTS
3. HHHHHHHHHHHHHCCCCCEEEEECCCCCCCCCCCCCCCCCCHHHHHHHCCCCCCCEEECCCCCCCCCCCCCCCCCCCCCC
4.
5.                                                      MIVGGHSQDPVCMAAENKKQVDYVPGT

1. PCKPDQQNGIWIVQAHEWGKYVGRADFEFRNGEMKMVNYQLIPVNLKKKVTWEDGKSERVLYTPEIAENQQMISLLS
2.     EEETTEEEE B STTSEEEEEEEEEETTEEEEEEEEEESS EEEEE SSS EEEEESS      HHHHHHHH
3. CCCCCCCCCCEEEEECCCCCCEEEEEEEEEECCEEEEEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHH
4.          IWIVQAHEWGKYVGRADFEFR β$*
5. PCKPDQQNGIWIVQ      YVGRADFEFRNGEMKMVNYQLIPVN*
```

Fig. 4 A testing sample [PDB: protein 1OID (A)] of the sequence level in the testing set.

Table 5 The predictive results of β-hairpins with ligand binding sites for 5-fold cross-validation and independent test.

	Acc (%)	*Mcc*	S_nH (%)	S_nNH (%)	S_pH (%)	S_pNH (%)
Training dataset	85.82	0.71	82.09	88.66	84.79	86.53
Testing dataset	84.78	0.70	85.39	86.05	81.13	89.34

obtained by predicted secondary structure from PSIPRED software, and the number of patterns with loop length of 2–12 was 341. Among the 341 patterns, 172 were assigned as β-hairpins by PROMOTIF software and the rest 169 patterns were assigned as non-β-hairpins. These data were used as independent testing dataset. The accuracy was 70.67% with Mcc of 0.41, and the sensitivity and the specificity for β-hairpin were 73.07% and 66.27%, respectively (Table 4). A sample (PDB: protein 1OID (A)) was given to explain the two different testing data (Fig. 4). It was obvious that the testing results of β-hairpins assigned by DSSP were better than those of β-hairpins assigned by PSIPRED. The reason behind this may be that DSSP can give the secondary structure more accurately, and this lays a foundation for the predictive process. Consequently, the predicted accuracy of ECE patterns based on better prediction of secondary structure was related to the prediction accuracy of β-hairpins directly. If the performance of the prediction of secondary structure can be improved, the prediction of β-hairpins will gain better results.

Note: The first three rows are amino acid sequence, observed secondary structure from DSSP and predicted secondary structure from PSIPRED, respectively. The other rows are ECE pattern predicted by PSIPRED; symbols of β, #, $ and * denote the β-hairpin assigned by PROMOTIF, the exact match, non-exact match, the correctly predicted β-hairpin and non-β-hairpin by our method, respectively.

3.3. Prediction for β-hairpins in enzymes with ligand binding sites

Furthermore, 245 features were input into SVM to predict β-hairpins with ligand binding sites: 845 β-hairpins with ligand binding sites and 1098 non-β-hairpins were randomly divided into training dataset (563 β-hairpins and 732 non-β-hairpins) and testing dataset (282 β-hairpins and 366 non-β-hairpins). The predicted results on training dataset (5-fold cross-validation) and testing dataset (independent test) are shown in Table 5.

It was shown that with 5-fold cross-validation on training dataset, the accuracy was 85.82% (Mcc of 0.71), and the sensitivity and the specificity for β-hairpin were 82.09% and 84.79%, respectively. For testing dataset in an independent test, the accuracy was 84.78% (Mcc of 0.70), and the sensitivity and the specificity for β-hairpin were 85.39% and 81.13%, respectively. Because the ligand binding site was crucial for activation of enzymatic reaction, the work will have important guiding significance for the experimental study of enzymes structure and function.

So far, the researches on enzyme mostly focus on the classification between enzyme and the non-enzyme (Cristian et al., 2008), and the identification of enzyme subclasses (Cai and Chou, 2005; Shi and Hu, 2010). There have been no reports about identification of the β-hairpin motifs in enzymes. In this work, taking into account the specific properties of β-hairpins in enzymes, we extracted the sequence information and predicted secondary structure information. Based the combined features, we adopted SVM algorithm in the prediction of β-hairpins in enzymes. The reasonable high prediction accuracy indicates that our method can be a valid tool for the further studies of β-hairpins in enzymes structure. What's more, this paper predicted β-hairpins with ligand binding sites, which was also a novelty for function prediction of enzymes. During the prediction process, we used mRMR criterion to filter features for the large number of original features and much

redundant information among the features that may bring problem of over fitting.

4. Conclusion

In this work, we constructed a dataset for β-hairpin in enzyme proteins from ArchDB_EC database, and β-hairpins containing ligand binding site also were given. We then constructed a testing dataset from ENZYME database that was completely irrelevant with ArchDB_EC database. For feature extraction, we only used sequence information and predicted secondary structure information. In case of over fitting, we used mRMR to optimize feature and reduce dimension. Some better results were obtained when feature optimization-based support vector machine method was used to recognize the β-hairpin motifs in enzymes.

In our future work, the comprehensive factors that facilitate the formation of β-hairpin motifs in enzymes are still need to investigate and used for the further prediction. Optimal dataset including more abundant experimental samples would be conducted, and extracting more relative biological features and using more valid algorithms would be our efforts to recognize the β-hairpin motifs in enzymes.

Web server

For facilitating study for other researchers, we developed an online web server. Based on our method, Apache and CGI-Perl 5.14.2 script as the background software were used to predict β-hairpin Motifs online, which is available at http://202.207.29.251:8080/. The predicted result was presented in table form and denotes which segment are the β-hairpins or non-β-hairpins.

Acknowledgments

This work was supported by National Natural Science Foundation of China (51467015 and 31260203) and Natural Science Foundation of the Inner Mongolia of China (2016MS0378).

References

Bonet, J., Planasiglesias, J., Garciagarcia, J., et al, 2014. ArchDB: structural classify cation of loops in proteins. Nucleic Acids Res. 42 (Database issue), D315–D319.

Cai, Y.D., Chou, K.C., 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J. Proteome Res. 4, 967–971.

Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 227, 45765–45769.

Cristian, R.M., Humberto, G.D., Alexandre, L.M., 2008. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J. Theor. Biol. 254, 476–482.

Cruz, X., Hutchinson, E.G., Shepherd, A., Thornton, J.M., 2002. Toward predicting protein topology: an approach to identifying β-hairpins. In: Proceedings of the National Academy of Sciences of the United States of America, USA, Aug. 20, pp. 11157–11162.

Ding, C.H.Q., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.

Edgar, J.R., Bell, R.M., 1978. Biosynthesis in *Escherichia coli* of sn-glycerol 3-phosphate, a precursor of phospholipid. J. Biol. Chem. 253, 6348–6353.

Espadaler, J., Fuentes, N.F., Hermoso, A., Querol, E., 2004. ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res. 32, 185–188.

Feng, Y.E., Kou, G.S., 2015. Identify beta-hairpin motifs with quadratic discriminant algorithm based on the chemical shifts. PLoS One 10 (9).

Fruton, J.S., 1999. Proteins, Enzymes, Genes——The Interplay of Chemistry and Biology. Yale University Press, New Haven.

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

Hu, X.Z., Li, Q.Z., 2008b. Using support vector machine to predict β-turns and γ-turns in proteins. J. Comput. Chem. 29, 1867–1875.

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.

Hutchinson, E.G., Thornton, J.M., 1996. PROMOTIF-a program to identify and analyze structural motifs in proteins. Protein Sci. 5, 212–220.

Jenny, T.F., Gerloff, D.L., Cohen, M.A., Benner, S.A., 1995. Predicted secondary and supersecondary structure for the serine–threonine-specific protein phosphatase family. Proteins 21, 1–10.

Jia, S.C., Hu, X.Z., Sun, L.X., 2013. The comparison between random forest and support vector machine algorithm for predicting β-hairpin motifs in proteins. Engineering 5, 391–395.

Kuhn, M., Meiler, J., Baker, D., 2004. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins 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. Nucleic Acids Res. 33, 154–159.

Lee, D.H., Jin, S.G., Cai, S., Chen, Y., et al, 2005. Repair of methylation damage in DNA and RNA by mammalian AlkB homologues. J. Biol. Chem. 280, 39448–39459.

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.

McGuffin, L.J., Bryson1, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.

Oliva, B., Bates, P.A., Querol, E., Aviles, F.X., et al, 1997. An automated classification of the structure of protein loops. J. Mol. Biol. 266, 814–830.

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 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 (August), 1226–1238.

Sander, C., Kabasch, W., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2637–2667.

Shi, R.J., Hu, X.Z., 2010. Predicting enzyme subclasses by using support vector machine with composite vectors. Protein Peptide Lett. 17, 599–604.

Shi, J.Y., Pan, Z., Zhang, S.W., Liang, Y., 2006. Protein fold recognition with support vector machines fusion network. Prog. Biochem. Biophys. 33, 155–162.

Stryer, L., Berg, J.M., Tymoczko, J.L., 2011. Biochemistry. W. H. Freeman, San Francisco.

Tatusova, T.A., Madden, T.L., 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 177, 187–188.

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

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience.

Wintjens, R.T., Rooman, M.J., Wodak, S.J., 1996. Automatic classification and analysis of alpha alpha-turn motifs in proteins. J. Mol. Biol. 255, 235–253.