Amino Acids (2010) 38:915–921
DOI 10.1007/s00726-009-0299-7

ORIGINAL ARTICLE

Recognition of β-hairpin motifs in proteins by using the composite vector

Xiu-Zhen Hu · Qian-Zhong Li · Chun-Lian Wang

Received: 10 December 2008 / Accepted: 20 April 2009 / Published online: 6 May 2009
© Springer-Verlag 2009

Abstract A composite vector method for predicting β-hairpin motifs in proteins is proposed by combining the score of matrix, increment of diversity, the value of distance and auto-correlation information to express the information of sequence. The prediction is based on analysis of data from 3,088 non-homologous protein chains including 6,035 β-hairpin motifs and 2,738 non-β-hairpin motifs. The overall accuracy of prediction and Matthew's correlation coefficient are 83.1% and 0.59, respectively. In addition, by using the same methods, the accuracy of 80.7% and Matthew's correlation coefficient of 0.61 are obtained for other dataset with 2,878 non-homologous protein chains, which contains 4,884 β-hairpin motifs and 4,310 non-β-hairpin motifs. Better results are also obtained in the prediction of the β-hairpin motifs of proteins by analysis of the CASP6 dataset.

Keywords β-Hairpin · Scoring matrix · Increment of diversity · Auto-correlation function

Introduction

Protein supersecondary structure prediction is a key step for predicting tertiary structure of proteins. The linker regions (linker region is also called loop) connect regular secondary structure α-helices and β-strands. In the β-strand–loop–β-strand motif, if the adjacent antiparallel β-strands are connected by one or more hydrogen bonds, the loop is called β-hairpin. Otherwise, the loop is non-β-hairpin (Kuhn et al. 2004). β-Hairpin is one of the frequently occurring motifs in proteins, and plays an important role in folding stability (Rose et al. 1985; Takano et al. 2000), recognition (Cruz and Thornton 1999; Rost et al. 1997) and structure assembly (Jones 2001) of a protein. When several adjacent antiparallel β-strands exist in proteins, the accurate prediction of β-hairpins can reduce the number of possible folding forms. Therefore, the prediction of β-hairpins is very important. An improvement in β-hairpin prediction will be helpful to molecular recognition studies and the prediction of protein tertiary structures.

Cruz et al. (2002) described an approach of predicting β-hairpins from predicted secondary structures. They calculated 14 scoring terms from the alignment. These scoring terms were used as the input units of a neural network for discriminating the potential β-hairpins and non-β-hairpins. The predictive accuracy was 47.7%. Kuhn et al. (2004) attempted to classify strand–loop–strand motifs by identifying local hairpins and non-local diverging turns using amino acid sequences. They achieved an accuracy of 77.3% in predicting hairpins (Kuhn et al. 2004). Recently, Kumar et al. predicted β-hairpins and non-β-hairpins in 2,880 non-redundant proteins with identity less than 33% using the single sequence information, evolutionary profile, surface accessibility and secondary structure information, a support vector machine (SVM) and artificial neural network (ANN) technique. They obtained an overall accuracy of 79.2% and Matthew's correlation coefficient of 0.59 (Kumar et al. 2005). Our previous works also proposed the SVM method for predicting β-hairpins with the loop length of 2–8 amino acids, based on the increment of diversity and

X.-Z. Hu · C.-L. Wang
College of Sciences, Inner Mongolia University of Technology, Hohhot 010051, People's Republic of China

Q.-Z. Li (✉)
Laboratory of Theoretical Biophysics, College of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, People's Republic of China
e-mail: qzli@imu.edu.cn

scoring function to express the information of sequence (Hu and Li 2008). A higher accuracy and Matthew's correlation coefficient were obtained from applying the method to the above two datasets ArchDB40 dataset as well as Kumar's dataset.

In this paper, the β-hairpin and non-β-hairpin with the loop length of 2–10 amino acids are selected, the bigger datasets than datasets in our previous works (Hu and Li 2008) are obtained. In order to compare with Kumar's choose method (Kumar et al. 2005), the same method for selecting fixed-length patterns contained 17 amino acids are used, which is different from our previous work. In addition, amino acid composition, dipeptide compositions from 20 amino acids and 6 hydropathy characteristics and surface accessibility are chosen as information parameters of the protein sequences. A composite vector including the score of matrix, increment of diversity, the value of distance and auto-correlation information for predicting β-hairpin motifs with the loop length of 2–10 amino acids is proposed. The higher accuracy and Matthew's correlation coefficient are obtained by using the new method.

Materials

The dataset from ArchDB40 (http://www.sbi.imim.es/cgi-bin/archdb/loops.pl) database includes 3,088 non-homologous protein chains with sequence identity <40%, resolution <3 Å, and 6,216 β-hairpin motifs and 2,964 non-β-hairpin motifs (Oliva et al. 1997; Espadaler et al. 2004).

The EVA dataset described in the work of Kumar et al. (2005) has 2,878 non-homologous protein chains with sequence identity <33%. The 5,437 β-hairpin motifs and 4,504 non-β-hairpin motifs are included in the EVA dataset. The secondary structure was assigned to each amino acid of proteins in the datasets using DSSP (Kabsch and Sander 1983). The program PROMOTIF (Hutchinson and Thornton 1996) was used to identify the observed β-hairpin and non-β-hairpin motifs.

The steps for generating the β-hairpin and non-hairpin datasets are similar to our previous method for predicting β-hairpins with the loop length of 2–8 amino acids (Hu and Li 2008). According to Kumar's work (Kumar et al. 2005), the motifs with the loop length of 2–10 amino acids are extracted. There are 6,035 β-hairpins and 2,738 non-β-hairpins in the ArchDB40, composing of the 98.2 and 97.1% of their corresponding total motifs. In EVA dataset, 4,884 β-hairpins and 4,310 non-β-hairpins are obtained, equal to 89.8 and 95.7% of their corresponding total motifs.

Methods

Position weight matrix (PWM)

In order to consider the effect of amino acids position conservation in β-hairpin sequence segments, PWM method is used. The PWM method had been widely used in the predicting of transcription factor binding sites in genomes (Wasserman and Sandelin 2004; Kielbasa et al. 2005; Kel et al. 2003). Each element in the PWM defined as:

$$w_{ij} = \log \frac{p_{ij}}{P_{0j}} \tag{1}$$

where P_{0j} is random probability and equals to 1/21, P_{ij} represents the probability of amino acids j at position i, which is defined as following (Wasserman and Sandelin 2004; Kielbasa et al. 2005):

$$P_{ij} = \frac{n_{ij} + \sqrt{N_i}/21}{N_i + \sqrt{N_i}} \tag{2}$$

here n_{ij} denotes the real counts for amino acid j (20 amino acid and one terminal residues) at ith position of the sequence segments; N_i is the total number of the sequences.

The PWM includes $21 \times L$ elements (L is length of the sequence segments). For an arbitrary sequence segment S with L amino acids (i.e., $S = (x_1, x_2, \ldots, x_L)$, where x_i is the amino acid at position i in segment S). The score of segment S can be defined as:

$$F(S) = F(x_1, x_2, \ldots x_L) = \sum_{i=1}^{L} w_{ij} \tag{3}$$

The class of sequence segment S may be predicted by the maximum among $F(S)^{\beta}$ and $F(S)^{\text{non-}\beta}$, and can be formulated as follows:

$$F(S)^{\xi} = \text{Max}\left\{F(S)^{\beta}, F(S)^{\text{non}-\beta}\right\} \tag{4}$$

here $\xi \in$ β-hairpin or non-β-hairpin. The operator Max means taking the maximum value among those in the parentheses, then the ξ will give the segment class to which the predicted segment should belong.

Fixed-length patterns of 17 amino acids were generated using the methods described below:

1. The loop locates the central position of the fixed-length pattern.
2. If pattern length was <17, residues flanking the peptide in the primary amino acid sequence were appended at both the ends.
3. If loop length was even, then nine residues from the left-hand side and eight residues from the right-hand side were taken.

Increment of diversity (ID)

Increment of diversity algorithm is a measure of composition similarity level for two systems (Laxton 1978). The ID algorithm has been applied in the recognition of protein structural class (Li and Lu 2001), the exon–intron splice site prediction (Zhang and Luo 2003) and prediction of subcellular location of proteins (Chen and Li 2007). In the state space of t dimension, the diversity measure for diversity source S: $\{m_1, m_2,\ldots, m_t\}$ is defined as (Laxton 1978; Li and Lu 2001):

$$D(S) = M \log M - \sum_{i}^{t} m_i \log m_i \tag{5}$$

In the same state space, the increment of diversity between the source of diversity $X(n_1, n_2,\ldots, n_t)$ and $S(m_1, m_2,\ldots, m_t)$ is defined as:

$$\mathrm{ID}(X, S) = D(X + S) - D(X) - D(S) \tag{6}$$

An arbitrary sequence segment may be predicted by the minimum among ID^{β} and $\mathrm{ID}^{\text{non-}\beta}$, and can be formulated as follows:

$$\mathrm{ID}^{\xi} = \mathrm{Min}\left\{\mathrm{ID}^{\beta}, \mathrm{ID}^{\mathrm{non}-\beta}\right\} \tag{7}$$

The meaning of symbol ξ is the same as Eq. 4. The operator Min means taking the minimum value among those in the parentheses, then the ξ will give the segment class to which the predicted segment should belong.

Distance method (DM)

If sequence segment S correspond to a M dimension vector $V = (f_1, f_2, \ldots, f_M)$, the similarity between S and S_i is defined (Chou and Cai 2006):

$$\Delta(V, V_i) = \frac{V \cdot V_i}{\|V\|\|V_i\|} \quad (i = \beta\text{-hairpin or non-}\beta\text{-hairpin}) \tag{8}$$

Generally speaking, the similarity is within the range of 0 and 1; i.e., $0 \le \triangle(V, V_i) \le 1$. Where $V{\cdot}V_i$ is the dot product of vectors V and V_i, the $\|V\|$ and $\|V_i\|$ are their magnitudes, respectively.The class of sequence segment S may be predicted by the maximum among $\triangle(V, V_{\beta})$ and $\triangle(V, V_{\text{non-}\beta})$, and can be formulated as follows:

$$\Delta(V, V_{\xi}) = \mathrm{Max}\left\{\Delta\left(V, V_{\beta}\right), \Delta\left(V, V_{\mathrm{non}-\beta}\right)\right\} \tag{9}$$

The meaning of symbol ξ is the same as Eq. 4. The ξ will give the segment class to which the predicted segment should belong.

Quadratic discriminant (QD)

The covariant discriminant function can be used as quadratic discriminant (QD), which was given by Chou (2000, 2005; Chou and Elrod 1998):

$$\mathrm{QD}_{\xi} = (x - \bar{x}_{\xi})^T \sum\nolimits_{\xi}^{-1} (x - \bar{x}_{\xi}) + \log\left|\sum\nolimits_{\xi}\right| \tag{10}$$

here $\bar{x}_{\xi}$ and Σ_{ξ} are the group mean and covariance matrix, respectively, (computed from the ξ training set), the recognition rule should be given by:

$$\mathrm{QD}_{\xi} = \mathrm{Min}\left\{\mathrm{QD}_{\beta}, \mathrm{QD}_{\mathrm{non}-\beta}\right\} \tag{11}$$

The meaning of symbol ξ is the same as Eq. 4. The ξ will give the segment class to which the predicted segment should belong.

Parameter selection

The probabilities of 21 amino acids (20 for the amino acid and 1 for terminal residues) at each position are selected as parameters (A_p) of PWM. By using of the training dataset to construct PWM (contained 21 × 17 elements), the two scores for every sequence segment can be obtained for the β-hairpins and non-β-hairpins.

In the ID algorithm, the frequencies of 400 dipeptide compositions (B_0) from 20 amino acids and 36 dipeptide compositions (C_0) from 6 hydropathy characteristics are selected as the parameters of diversity source of the β-hairpins and non-β-hairpins, respectively. They constructed 20 × 20 and 6 × 6 dimensions state space, respectively. The classifications of hydropathy characteristics for amino acids (Chen and Li 2007) are shown in Table 1.

The frequencies of 20 amino acids (A_d) are selected as parameters for the distance method (DM). Based on the Eq. 8, for every sequence segment, the two values of the distances can be obtained for the β-hairpins and non-β-hairpins.

When using QD algorithm to predict β-hairpins and non-β-hairpins, for every sequence segment, two scores can be obtained by the PWM algorithm, the four increment of diversity (for B_0 and C_0 parameters) values can be obtained by ID algorithm, and two values of the distances

Table 1 Hydropathy characteristics for amino acids

Classification	Amino acids	Classification	Amino acids
Strongly hydrophilic or polar	R, D, E, N, Q, K, H	Proline	P
Strongly hydrophobic	L, I, V, A, M, F	Glycine	G
Weakly hydrophilic or weakly hydrophobic	S, T, Y, W	Cysteine	C

can be obtained by DM algorithm. The eight parameters are selected as the inputting parameters of the QD.

Performance measures

In order to evaluate the correct prediction rate and the reliability of a predictive method, the accuracy (Acc), Matthew's correlation coefficients (Mcc), β-hairpin sensitivities ($Q_{o(\mathrm{H})}$), non-β-hairpin sensitivities ($Q_{o(\mathrm{NH})}$), β-hairpin specificities ($Q_{p(\mathrm{H})}$) and the non-β-hairpin specificities ($Q_{p(\mathrm{NH})}$) are calculated by:

$$\mathrm{Acc} = \frac{(p+r)}{(p+r+o+u)} \times 100 \tag{12}$$

$$\mathrm{Mcc} = \frac{(p \times r) - (o \times u)}{\sqrt{(p+u)(p+o)(r+u)(r+o)}} \tag{13}$$

$$Q_{o(\mathrm{H})} = \frac{p}{p+u} \times 100 \tag{14}$$

$$Q_{o(\mathrm{NH})} = \frac{r}{r+o} \times 100 \tag{15}$$

$$Q_{p(\mathrm{H})} = \frac{p}{p+o} \times 100 \tag{16}$$

$$Q_{p(\mathrm{NH})} = \frac{r}{r+u} \times 100 \tag{17}$$

There, p and r are, respectively, the numbers of correctly predicted β-hairpin and non-β-hairpin motifs, u is the number of β-hairpin motifs but is predicted as non-β-hairpin motifs and o is the number of non-β-hairpin motifs that is missed by the prediction.

Results and discussion

Prediction using fivefold cross-validation for β-hairpins in the ArchDB40 dataset

The predictive results by using PWM algorithm

The predicting results of β-hairpins and non-β-hairpins are shown in Table 2. The Mcc and Acc values are 0.39 and 70.1%, respectively.

The predictive results by using ID algorithm

For each type parameter (B_0 and C_0), the performance of ID method is also shown in Table 2. Similar results are obtained by the two types of parameters. The Mcc are 0.22 and 0.19, respectively. And they are lower than 0.39 from the PWM method.

The predictive results by using DM algorithm

The performance of distance method is also shown in Table 2. Similar results are obtained by the ID method; the Mcc and Acc are 0.22 and 60.9%, respectively.

The predictive results by using QD algorithm

The above predicted results of the β-hairpins demonstrate that an algorithm with single parameter may provide partial information of a sequence. Therefore, to enhance the prediction performance, a composite vector integrating the above calculated PWM (A_p), ID (B_o), ID (C_o) and DM (A_d) is selected as the inputting parameters of the QD. The results are shown in Table 2. The Mcc value is increased to 0.44, and overall prediction accuracy is increased to 76.4%. These results indicate the prediction is improved by using the new method.

Compared to the results from our previous work (Hu and Li 2008) by using ArchDB40 dataset the results from the new method are slightly lower. Acc and Mcc in our previous work (Hu and Li 2008) are 79.9% and 0.59, respectively. But the dataset in this paper is bigger than previous dataset (Hu and Li 2008) and the loop length contains more amino acids.

According to Kumar's work (Kumar et al. 2005), the surface accessibility value for residue is impotent parameters, and auto-correlation function can reflect information of the protein structure (Kawashima et al. 1999; Zhang et al. 2004). In order to further improve predictive effect, a new auto-correlation function Z score is added into the composite vector. For an arbitrary sequence segment S with L amino acids (i.e., $S = (x_1, x_2, \ldots, x_L)$, auto-correlation function Z can be defined as follows:

Table 2 The performance of various methods in the ArchDB40 dataset

Method (parameters)	$Q_{o(\mathrm{H})}$	$Q_{o(\mathrm{NH})}$	$Q_{p(\mathrm{H})}$	$Q_{p(\mathrm{NH})}$	Acc	Mcc
PWM (A_p)	68.6	73.5	85.5	50.7	70.1	0.39
ID (B_o)	62.3	61.0	78.5	41.5	61.9	0.22
ID (C_o)	63.4	57.0	77.1	40.6	61.5	0.19
D (A_d)	59.5	64.0	79.0	40.9	60.9	0.22
QD [PWM (A_p), ID (B_o), ID (C_o), D (A_d)]	83.0	61.4	83.1	61.3	76.4	0.44
QD [PWM (A_p), ID (B_o), ID (C_o), D(A_d), Z]	91.3	64.3	85.4	76.4	83.1	0.59

Table 3 Surface accessibility value of amino acid residue

Amino acid	A	R	N	D	C	Q	E	G	H	I
Accessible surface area	25	90	63	50	19	71	49	23	43	18
Amino acid	L	K	M	F	P	S	T	W	Y	V
Accessible surface area	23	97	31	24	50	44	47	32	60	18

$$Z = \frac{1}{L-1}\sum_{j=1}^{L-1} h_j h_{j+1} \tag{18}$$

where h_j denotes the surface accessibility value for residue x_j. The surface accessibility values of 20 amino acids are shown in Table 3. (http://www.genome.ad.jp/dbget/aaindex.html)

The PWM (A_p), ID (B_o), ID (C_o), DM (A_d) values and surface accessibility auto-correlation information are all selected as inputting parameters for QD. The performance is also shown in Table 2. The $Q_{o(\mathrm{H})}$ value is 91.3%, $Q_{o(\mathrm{NH})}$ is 64.3%, $Q_{p(\mathrm{H})}$ value is 85.5%, $Q_{p(\mathrm{NH})}$ is 76.4%, Acc is 83.1% and Mcc is 0.59.

The new results of predicting β-hairpins are better than the prediction from our previous work (Hu and Li 2008). When the auto-correlation information is added into input parameters, the prediction accuracy is gained about 6%.

Prediction using fivefold cross-validation for β-hairpins in the Kumar's dataset

The above methods are also applied to Kumar's dataset in predicting β-hairpins with fivefold cross-validation. The predictive results by using PWM, ID, DM and QD algorithm are shown in Table 4. In order to compare our method with other methods, Kumar's results from the same dataset with fivefold cross-validation is also shown in Table 4.

Predicting results in the Table 4 indicate that the composite vector containing PWM (A_p), ID (B_o), ID (C_o), DM (A_d) and Z values is used as the input parameters of the QD, the best predictive results are obtained. In addition, the auto-correlation function Z is also important informational parameter.

Prediction for β-hairpins in the CASP6 proteins

In order to further evaluate the new predictive method, the β-hairpins and non-β-hairpins in the CASP6 proteins as an independent testing dataset of the EVA dataset are predicted using our new method. The dataset contains 78 β-hairpins and 102 non-β-hairpins (the lengths of loops are 2–10 amino acids). When PWM (A_p), ID (B_o), ID (C_o), DM (A_d) and Z are selected as the inputting parameters of QD algorithm, the predictive results are shown in Table 5. Kumar's results (Kumar et al. 2005) from the prediction of the same dataset are also shown in Table 5. For more detail comparison between two methods, the prediction result for 78 β-hairpins in the CASP6 dataset are shown in Table 6.

Conclusion

The successful prediction indicates that using the scores of sequence segment, the increments of diversity, the values of distance and auto-correlation information as the parameters

Table 4 The performance of various methods in the EVA (Kumar's) dataset

Method (parameters)	$Q_{o(\mathrm{H})}$	$Q_{o(\mathrm{NH})}$	$Q_{p(\mathrm{H})}$	$Qp_{(\mathrm{NH})}$	Acc	Mcc
PWM (A_p)	67.2	64.5	69.8	61.8	66.0	0.32
ID (B_o)	61.3	58.5	64.2	55.4	60.2	0.20
ID (C_o)	64.1	58.3	65.2	57.2	61.5	0.22
D (A_d)	65.8	52.2	62.6	55.6	59.6	0.18
QD [PWM (A_p), ID (B_o), ID (C_o), D (A_d)]	81.3	68.0	75.6	74.9	75.3	0.50
QD [PWM (A_p), ID (B_o), ID (C_o), D (A_d), Z]	83.4	77.4	81.8	79.3	80.7	0.61
Kumar et al's evolutionary profile	82.6	75.7	77.2	81.4	79.2	0.56

Table 5 The predicted results of β-hairpins and non-β-hairpins in the CASP6 Dataset

	Loop (2-10)	Hairpin motifs	Exact matches	Non-exact matches	Non-hairpin motifs
Our method Acc = 75.6%	Number	78 (55)	27 (23)	51 (32)	102 (81)
	Q_o	70.5%	85.2%	62.7%	79.4%
Kumar's method Acc = 73.3%	Number	78 (47)	27 (22)	51 (25)	102 (85)
	Q_o	60.3%	81.5%	49.0%	83.3%

Number in the parentheses is the number of the correctly predicted motifs; results of Kumar's method are from (Kumar et al. 2005)

Table 6 Comparison of predictive results for β-hairpins of 63 proteins in the CASP6 dataset

β-hairpins	S1	S2	β-hairpins	S1	S2
β-hairpins of exact matches					
RWVYKLNGVTLEVNRV	True	True	CLIVEIGGVYFVRR	True	True
EKFVLENGVL	False	True	TAIVQIRNREMPVKVT	True	True
LGIVSGGRLI	True	True	RYELRNGEIRAT	True	True
ISGEFSLFAKGYWVENGEIA	True	True	FRIHAIAGGYRFLT	True	True
NIVIKLLEVNGNHAIKIS	True	True	KMIDVALRVDGVEVDRIR	True	True
TREFSLRLANGDLDQYTD	True	True	MIQMGTKFYQI	True	True
LTIQVNGVP	True	True	VIDESSHFVSVA	False	True
LNGVALHFEGGKATAAERFI	True	True	KAIVVADGQKSV	False	False
TDLAVLSDGELQLTTI	True	False	KVELRQVECINGNIFLHLGAV	True	True
SIHVDGEGTCLVT	True	False	MMFHVRTDSNHDVLMD	True	True
VKTLIVLDNAGGVYAVVI	True	True	KFYQIDSTGKLSE	True	True
YGTYGMVSESGEHFNAI	True	True	RLFVAESESGEVVGFAAF	True	True
GQVKVKFDVTPDGRVDNVQILS	True	False	HYQAFIRDELDENKWYKF	True	True
LLWRLEPARGLEPPVVLVQ	False	False			
The β-hairpins of non-exact matches					
IKVTVTNSFFEV	False	False	VGFKVKGPSGIGGIVRI	False	False
RWIGNMMFHVR	True	False	LGLVYDIQIDDQNNVKVLMTM	True	True
KVERMSKTVYTVDWV	True	True	PVRFTDAQGNQHEGIIT	True	True
REIELGGAKLWEVAYGF	True	True	VICKPIGPSKVYVS	False	False
IVYYFTEDFFRLVV	True	True	RLLEGERGPWVQI	True	False
FNVEIKVLKDGKTGTFRT	True	True	FVSVAPFAATYPFEIWI	False	False
VLRVGRFEDDGYFCTIEVTATSTVT	True	True	RMVDFHGWMMPLHY	False	False
VISFEGGKLKVRVKA	True	True	DLFIATTGYTGEAGYEIAL	True	True
YSYKYVHDDGRVSYPLCFIF	True	True	TFSPTLGYSIALARV	False	False
FVIDDAKNIYIYVSG	True	True	VNVLFVDDEAKTNQIFARRRLSFDCTATLK	True	False
WYKFNDDKVSVV	True	True	KDFRIEIEYERTEEHPRIFTKVHLKYIFKF	True	False
GWRTFDVNGEKLTVVNLL	True	True	YNLIGVITHQGANSESGHYQAFIR	False	False
VYFEVPRPKLLRIR	False	False	VVLVRKVGAPGNPEFALGAV	False	False
FAVFSGKYFKGESPIGSVYLF	True	True	FLFAMAIARDANPRSGSWYELAR	True	True
KIKYTGGELCI	False	True	HLHFITEDKTSGGHVLNLQFD	True	True
TVRGVFIVDARGVIRTMLYY	True	True	LIFSILTEFGVSKVTPIK	True	True
AIIMKVDKDRQMVVL	True	True	FVHFRNVKYLGEHRFE	False	True
ILRVMLIPVSEDYLIFSIL	False	True	HWIITDANGKTSEVQ	True	True
GGIVRIERNREKVEFAI	True	True	RILKGGTAYQT	True	False
YVNFYIANGGII	True	False	MCCFARPGVVLLSW	False	False
RLYTHPDGRIVVVP	False	False	QVEYFNSKLKQKFTLTLG	True	True
LAFDREGYRL	True	False	MEVTTDHGVIKL	False	False
LEVNRVEGIGDFVDIEV	False	False	KALYSGMLNASGGVID	True	False
CFILIPIAPFRFGWKPYVV	False	False	SIRVEVKTEYIEQQSSPEDEKYLFSYTITIIN	True	False
GTCWMARSPHPDVPYIIVF	False	False	FLWKVWTESEKNHEAGGIYLFT	True	False
GLYVLTAKDGDEVAAGTVNWVT	False	False			

True shows correct prediction, *False* shows incorrect prediction, *S1* results by our method, *S2* results by Kumar's method (Kumar et al. 2005)

of QD can reduce the dimension of input vector, improve calculating efficiency and extract important classify information. It is possible that the QD algorithm plays a role of syncretizing information. Moreover, the auto-correlation information of surface accessibility is a useful parameter for improving predictive performance of the β-hairpins.

Acknowledgments This work was supported by National Natural Science Foundation of China (30560039) and Project for Excellent Subject-directors of Inner Mongolia Autonomous Region.

References

Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783. doi:10.1016/j.jtbi.2006.11.010

Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278:477–483. doi:10.1006/bbrc.2000.3815

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19. doi:10.1093/bioinformatics/bth466

Chou KC, Cai YD (2006) Prediction of protease types in a hybridization space. Biochem Biophys Res Commun 339:1015–1020. doi:10.1016/j.bbrc.2005.10.196

Chou KC, Elrod DW (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 252:63–68. doi:10.1006/bbrc.1998.9498

Cruz X, Thornton JM (1999) Factors limiting the performance of prediction-based fold recognition methods. Protein Sci 8:750–759

Cruz X, Hutchinson EG, Shepherd A, Thornton JM (2002) Toward predicting protein topology: an approach to identifying β-hairpins. Proc Natl Acad Sci USA 99:11157–11162. doi:10.1073/pnas.162376199

Espadaler J, Fuentes NF, Hermoso A, Querol E, Aviles FX, Sternberg MJE, Oliva B (2004) ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res 32:185–188. doi:10.1093/nar/gkh002

Hu XZ, Li QZ (2008) Prediction of the β-hairpins in proteins using support vector machine. Protein J 27:115–122. doi:10.1007/s10930-007-9114-z

Hutchinson EG, Thornton JM (1996) PROMOTIF-A program to identify and analyze structural motifs in proteins. Protein Sci 5:212–220

Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. Proteins 5:127–132. doi:10.1002/prot.1171

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637. doi:10.1002/bip.360221211

Kawashima S, Ogata H, Kanehisa M (1999) Aaindex: amino acid index database. Nucleic Acids Res 27:368–369. doi:10.1093/nar/27.1.368

Kel AE, Goßling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31:3576–3579. doi:10.1093/nar/gkg585

Kielbasa SM, Gonze D, Herzel H (2005) Measuring similarities between transcription factor binding sites. BMC Bioinformatics 6:237. doi:10.1186/1471-2105-6-237

Kuhn M, Meile J, Baker D (2004) Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Bioinformatics 54:282–288

Kumar M, Bhasin M, Natt NK, Raghava GPS (2005) BhairPred: prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33:154–159. doi:10.1093/nar/gki588

Laxton RR (1978) The measure of diversity. J Theor Biol 71:51–67. doi:10.1016/0022-5193(78)90302-8

Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213:493–502. doi:10.1006/jtbi.2001.2441

Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE (1997) An automated classification of the structure of protein loops. J Mol Biol 266:814–830. doi:10.1006/jmbi.1996.0819

Rose GD, Gierasch L, Smith JA (1985) Turns in peptides and proteins. Adv Protein Chem 37:1–109. doi:10.1016/S0065-3233(08)60063-7

Rost B, Schneider R, Sander C (1997) Protein fold recognition by prediction-based threading. J Mol Biol 270:471–480. doi:10.1006/jmbi.1997.1101

Takano K, Yamagata Y, Yutani K (2000) Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. Biochemistry 39:8655–8665. doi:10.1021/bi9928694

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276–287. doi:10.1038/nrg1315

Zhang LR, Luo LF (2003) Splice site prediction with quadratic discriminate analysis using diversity measure. Nucleic Acids Res 31:6214–6220. doi:10.1093/nar/gkg805

Zhang SW, Pan Q, Zhang HC, Wang HY, Zhang MG (2004) Prediction of multi-class protein folds by using support vector machine. J Northwest Polytech Univ 22(2):200–204