Amino Acids (2012) 42:1749–1755
DOI 10.1007/s00726-011-0889-z

ORIGINAL ARTICLE

Using predicted shape string to enhance the accuracy of γ-turn prediction

Yaojuan Zhu · Tonghua Li · Dapeng Li · Yun Zhang · Wenwei Xiong · Jiangming Sun · Zehui Tang · Guanyan Chen

Received: 17 December 2010 / Accepted: 9 March 2011 / Published online: 22 March 2011

Abstract Numerous methods for predicting γ-turns in proteins have been developed. However, the results they generally provided are not very good, with a Matthews correlation coefficient (MCC) $\leq$0.18. Here, an attempt has been made to develop a method to improve the accuracy of γ-turn prediction. First, we employ the geometric mean metric as optimal criterion to evaluate the performance of support vector machine for the highly imbalanced γ-turn dataset. This metric tries to maximize both the sensitivity and the specificity while keeping them balanced. Second, a predictor to generate protein shape string by structure alignment against the protein structure database has been designed and the predicted shape string is introduced as new variable for γ-turn prediction. Based on this perception, we have developed a new method for γ-turn prediction. After training and testing the benchmark dataset of 320 non-homologous protein chains using a fivefold cross-validation technique, the present method achieves excellent performance. The overall prediction accuracy Q_{total} can achieve 92.2% and the MCC is 0.38, which outperform the existing γ-turn prediction methods. Our results indicate that the protein shape string is useful for predicting protein tight turns and it is reasonable to use the dihedral angle information as a variable for machine learning to predict protein folding. The dataset used in this work and the software to generate predicted shape string from structure database can be obtained from anonymous ftp site ftp://cheminfo.tongji.edu.cn/GammaTurnPrediction/ freely.

Y. Zhu · T. Li (✉) · D. Li · Y. Zhang · W. Xiong · J. Sun ·
Z. Tang · G. Chen
Department of Chemistry, Tongji University,
Room 438, No.1239, Siping Road, Shanghai 200092,
People's Republic of China
e-mail: lith@tongji.edu.cn

Keywords Imbalanced data · Shape string · Support vector machine (SVM) · γ-Turn prediction

Introduction

Protein secondary structure is composed of regular elements, such as α-helices, and β-sheets, and irregular elements, such as tight turns, bugles, and random coils. A tight turn has been described as a site where (a) the polypeptide chain reverses its overall direction, which causes the chain to fold back on itself and (b) the number of amino acids directly involved in forming the turn is no more than six. According to the number of residues involved in forming the turns, tight turns can be categorized as δ-turns, γ-turns, β-turns, α-turns and π-turns (Chou 2000). They play a vital role in folding compact globular structures and molecular recognition because they usually occur on the exposed surface of proteins (Rose et al. 1985) and defining template structures used for the design of new molecules such as drugs, pesticides and antigens.

The γ-turn is the second most characterized and commonly found tight turn after the β-turn, which involves three amino acid residues and a hydrogen bond between the backbone $CO_{(i)}$ and the backbone $NH_{(i+2)}$. There are two types of γ-turns: inverse and classic (Bystrov et al. 1969). On average, γ-turns account for 3.4% of total amino acid contents of protein structures (Guruprasad and Rajkumar 2000). The problems of γ-turn prediction can be divided into two categories; prediction of γ-turn types (Chou 1997a; Chou and Blinn 1997; Jahandideh et al. 2007) and prediction of γ-turn/non- γ-turn (Guruprasad et al. 2003; Kaur and Raghava 2002; Pham et al. 2005). In the past, methods have been developed for the prediction of γ-turns based on the statistical model and machine learning technique (Alkorta et al. 1996; Kaur and Raghava 2002; Guruprasad et al. 2003). The statistical methods include the sequence coupled model and the GOR model (Chou 1997a, b; Chou and Blinn 1997; Garnier et al. 1978; Gibrat et al. 1987). The machine-learning techniques include the neural network (Zell and Mamier 1997; Wrtten and Frank 1999), SNNS, using multiple sequence alignment as input instead of single amino acid sequence (Kaur and Raghava 2003) and GTSVM, which employs SVM to predict γ-turns and yields a good performance (Pham et al. 2005). All of these methods greatly outperformed statistical approaches. However, the prediction performance is still restricted in comparison to predictors of other tight turns due to the complexity of the problem and the imbalanced nature of the data. The fact is that γ-turn consists of three residues and thus is more flexible than other tight turns. Moreover, the present dataset has a ratio of $\sim$30:1 of non-γ-turn and γ-turn residues.

In this work, we introduce a modified optimal criterion for SVM to overcome the challenge of the imbalanced problem in the training data and design a predictor of shape string based on structure alignment. The predicted shape string is considered a new variable for input to SVM for γ-turn prediction and the results show that these innovations significantly improve the performance of γ-turn prediction. The MCC is 0.38 and the overall prediction accuracy achieves 92.2% for 320 non-homologous protein chains in fivefold cross-validation.

Materials and methods

Dataset

The benchmark dataset of 320 non-homologous protein chains first described by Guruprasad and Rajkumar (2000) was chosen to train and test our method. The structure of each protein chain in this dataset is determined by X-ray crystallography at better than 2.0 Å resolution, and no two protein chains share more than 25% sequence identity. All observed γ-turns are identified by the PROMOTIF program (Hutchinson and Thornton 1996). Each chain contains at least one γ-turn.

Method

The flowchart of the present method for γ-turn prediction is shown in Fig. 1.

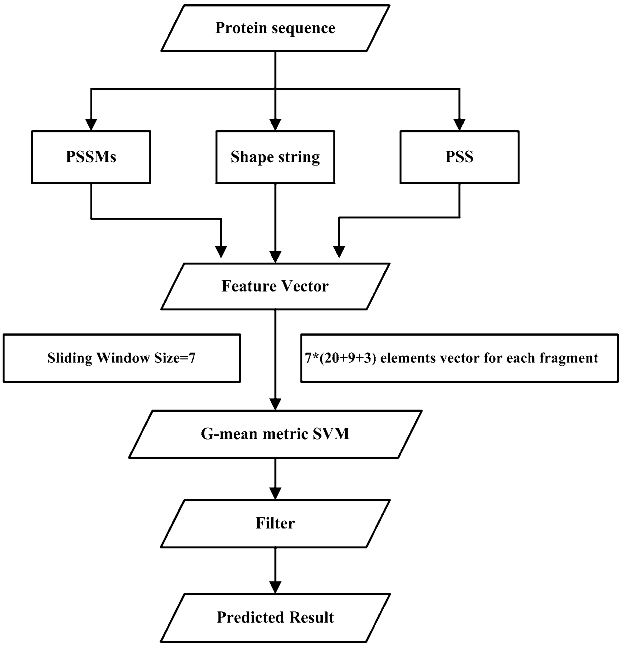

Fig. 1 Architecture of the present γ-turn prediction system. The major difference between this and previous approaches is addition of the shape string variable. Our experiments confirm that shape string plays a vital role in γ-turn prediction

A given protein sequence is represented by the position-specific scoring matrices (PSSMs), predicted secondary structure (PSS) and shape strings. PSSMs, which reflect the evolutionary information between a given protein and its remote homologs, are generated using the PSI-BLAST program (Altschul et al. 1997) and these profiles are scaled to a range of 0–1 using the standard logistic function. The predicted secondary structure of a given protein generated by PSIPRED (Jones 1999) is encoded as follows: helix→(1,0,0), strand→(0,1,0), coil→(0,0,1). Shape strings for a given protein are generated by a predictor that we designed and they are represented by nine characters (for detail see next section). Each protein is subsequently sliced into fragments of seven amino acids with sliding window. These 7*(20 + 9 + 3) elements are constituted as feature vectors of SVM for training and predicting. The output of SVM is filtered using "state-flipping" rule (Shepherd et al. 1999) and we can obtain the final predicted results.

Structure database preparation

Before considering shape string, we constructed a structure database (SD) containing protein sequences and their shape strings. SD_$_{100}$ is a non-redundant (nr) Protein Data Bank (PDB) database of 46860 entries, which are used as the nr-PDB database of the Basic Local Alignment Search Tool (BLAST) web server and can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db (accessed September of 2010), after removing the 320 non-homologous protein chains. Three other SD databases at different non-redundant levels, culling SD_$_{100}$ at 30% (SD_$_{30}$), 60% (SD_$_{60}$), 90% (SD_$_{90}$) sequence identity by using CD-HIT (Li and Godzik 2006) are generated for comparison. For each entry

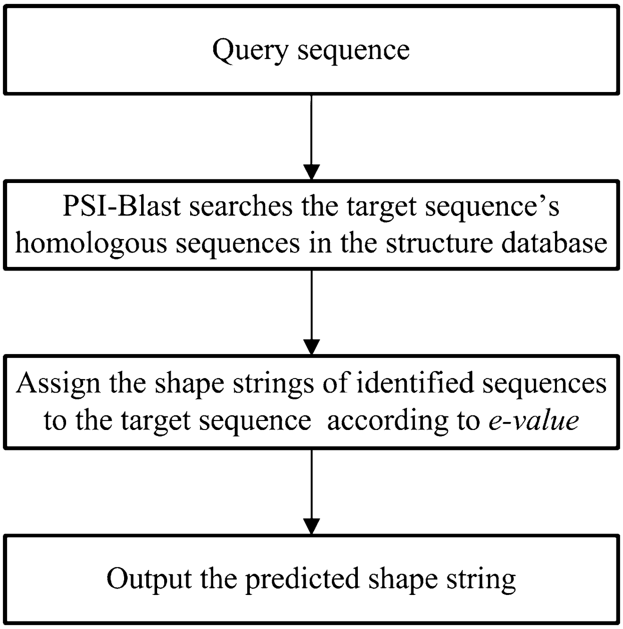
Query sequence
PSI-Blast searches the target sequence's homologous sequences in the structure database
Assign the shape strings of identified sequences to the target sequence according to *e-value*
Output the predicted shape string

Fig. 2 Outline of the predicted shape string generation procedure

in SD, its shape strings are obtained from the web server (Hovmöller and Zhou 2004) and stored correspondingly.

Shape string

Recently, new prediction approaches have been proposed based on structure alignment and they significantly promote the accuracy of predictors for protein secondary structure (DiFrancesco et al. 1996; Montgomerie et al. 2006). Dihedral angle of a protein backbone is also a characteristic of its secondary structure and is usually described by Φ/Ψ pair in the Ramachandran plot (Hovmöller et al. 2002) and yet can be expressed as shape string (Ison et al. 2005). There are eight characters (S, R, U, V, K, A, T and G), which are used to record shape strings. Shape string is only an expression of protein secondary structure but since it contains clustering information, we believe shape string will play an important role in the prediction of protein structure and function, especially for tight turns. Shape string has previously been used in the prediction of protein secondary structure (Zhou et al. 2010). For a protein of known structure, its shape strings can be calculated according to three-dimensional structures determined experimentally. For a protein of unknown structure, the key is to obtain its shape strings accurately. One can predict the shape string based on the available information. Here, we have designed a shape string predictor based on structure alignment (Fig. 2).

For a query or target sequence, the PSI-BLAST program is run on the SD database to find its homologous sequences. The identified sequences whose e-values are below a given threshold, for example 10^{-5} are next identified and ranked according to e-value in ascending order. The ranked sequences are then judged individually. When the top aligned sequence has a portion matching to the target sequence, the shape strings corresponding to this portion are assigned as the target's shape strings. Only those positions of the target sequence which have not been matched are left to judge the next aligned sequence (Fig. 3).

This assignment is repeated until all the aligned sequences are judged. If a query sequence has more regions homologous to the existing sequences in SD, the shape string of the target will be predicted more accurately. There are some empty positions of the query sequence where are never matched and have no shape string information. For an empty position, it is expressed as 'X', then shape string is represented by nine characters (S, R, U, V, K, A, T, G and X). In this study, nine characters are encoded by using the unary encoding scheme in which each of the characters is represented by a single one (which varies in position depending on the type) and eight zeros ($S = 1000000$, $R = 0100$, ($X = 00001$).

Query sequence: MKKVLITGF......KQIGNAM......NKFFLLGKN......VSLDYLEKDR......

Blasted result:

ID	e-value	MKKVLITGF......KQIGNAM......NKFFLLGKN......VSLDYLEKDR......
1IU8A	6e-064	--**SSSSRSS**......**AS**---**AK**......**KS**-----**RT**......**AAAAAAAAAK**......
2DF5A	2e-062	--SSSSSSS......RS**SGK**RR......AA**AK-KS**RS......AAA-------......
3LACA	9e-048	-**R**SSSSSSR......RSSGKSS......KU-----RT......AA--------......
predicted shape strings		**X**RSSSSRSS......ASSGKAK......KSAK**X**KSRT......AAAAAAAAAK......

Fig. 3 Detail illustrating how the shape string is generated by our predictor. The Blast result contains the identity of homologous sequences, the corresponding e-value, and the actual sequences. We show the corresponding shape string. *Bold letters* denote the shape strings assigned to the target sequence and they are merged to produce the final continuous result. Empty positions are denoted by *X* in *bold font* in the final result. A *hyphen* (-) means that there is no sequence homologous to the query sequence at this position. An *ellipsis* (…) denotes parts of the protein not shown in the figure

PSSMs and PSS

With the multiple alignments, we use the position-specific scoring matrices (Guruprasad et al. 2003; Kaur and Raghava 2003; Pham et al. 2005; Zhang et al. 2005) generated by the PSI-BLAST program searching against the large non-redundant database. These profiles are scaled to 0–1 using the standard logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{1}$$

where x is the raw profile matrices value. The predicted secondary structure from PSIPRED (Jones 1999) is also used.

Imbalanced problem

Using accuracy to evaluate the classifier on highly imbalanced datasets is not as good as it is on balanced datasets because the classifier will label all samples as the majority class to achieve the high predictive accuracy and fail on the minority class (Barandela et al. 2003). The geometric mean is a good indicator of the classifier performance in this condition because it is independent of the distribution of examples between positive and negative classes, and as Kubat et al. (1997) suggested, the geometric mean (G-mean) metric is defined as:

$$g = \sqrt{a^{+}a^{-}} \tag{2}$$

where a^{+} is the accuracy on the positive examples and a^{-} denotes the accuracy on the negative examples. This measure tries to maximize the accuracy of both classes while keeping the two accuracies balanced. Several researchers have used this metric for evaluating classifiers on imbalanced datasets (Kubat and Matwin 1997; Robert et al. 1997; Wu and Chang 2003; Anand et al. 2010). We also utilize this metric to evaluate SVM classifier for the high imbalanced γ-turn dataset and modify the evaluation criterion of LibSVM (Chang and Lin 2001) using the G-mean metric in this study.

Training and testing

We employ fivefold cross-validation to evaluate the performance of the present method. The 320 protein chains are randomly divided into five subsets, each containing approximately equal number of proteins as previously described (Zhang et al. 2005). The sliding window technique with seven residues is selected to get the best result and is applied to every protein in each subset. Each subset is an imbalanced set that remains the naturally occurring proportion of γ-turns and non-γ-turns. The method has been trained on four subsets, and the performance is measured by the remaining fifth set. This process is repeated five times so that each set is tested and an average accuracy is computed. The support vector machine (SVM), which is a popular algorithm from the machine learning community and has been widely applied in biology (Cai et al. 2002; Chou and Cai 2002) is chosen. The radial basis function is employed as kernel function of LibSVM (Chang and Lin 2001) and the weight factor is set to 30 according to a non-γ-turn/γ-turn ratio of $\sim$30:1.

Filtering

The prediction is performed separately for each residue and without reference to the prediction status of neighboring residues. Predictions include several unusually short γ-turns of one or two residues. To ensure that γ-turn is at least three residues long, we have added a simple filtering step known as the “state-flipping” rule as first described by Shepherd et al. (1999).

Performance measures

We adopt four criteria described by Shepherd et al. (1999) to evaluate the prediction performance of predictive method in accordance with previous papers on γ-turn prediction. (1) Q_{total} (prediction accuracy); the percentage of correctly predicted residues. (2) MCC; which accounts for both over- and under-prediction. (3) $Q_{\text{predicted}}$; the percentage of correct prediction of γ-turn residues (or probability of correct prediction). (4) Q_{observed}; the percentage of observed γ-turn residues that are correctly predicted (or percent coverage). These parameters can be calculated by the following equations:

$$Q_{\text{total}} = \left(\frac{p+n}{t}\right) \times 100 \tag{3}$$

$$\text{MCC} = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \tag{4}$$

$$Q_{\text{predicted}} = \left(\frac{p}{p+o}\right) \times 100 \tag{5}$$

$$Q_{\text{observed}} = \left(\frac{p}{p+u}\right) \times 100 \tag{6}$$

where p (number of correctly classified γ-turn residues), n (number of correctly classified non-γ-turn residues), o (number of non-γ-turn residues incorrectly classified as γ-turn residues), and u (number of γ-turn residues incorrectly classified as non-γ-turn residues). $t = p + n + o + u$ is the total number of residues.

Results

Prediction results using shape string variable

Results for the γ-turn/non-γ-turn prediction from the present method using shape string variable as input for training and testing are shown in Table 1. SSD stands for the shape strings generated by our predictor from structure database. For example, SSD$_{\_30}$ is the shape strings generated by our predictor searching against SD$_{\_30}$. SSD$_{\_60}$, SSD$_{\_90}$ and SSD$_{\_100}$ are obtained similarly. SSD$_{\_real}$ denotes the real shape strings obtained from the web server (Hovmöller and Zhou 2004).

It is observed that shape string is an extraordinary useful variable for γ-turn prediction and we can obtain a satisfactory prediction result. The MCC is over 0.30 and the Q_{total} exceeds 90% under different conditions. It is also evident that the shape string generated from different structure databases has a considerable effect on the result. This is because the sequence identity cut-off is at a higher level, that is, more homologous sequence fragments existing in the structure database, and a greater portion of the queried sequence can be matched by the sequences in the structure database; then the predicted shape string generated by our predictor are closer to the real. We can see that if we use real shape string as input we can obtain much better results. The MCC can achieve 0.58 and the Q_{total} is 96.0%. We can also see that shape string prediction accuracy increases with more homologous sequences existing in the structure database (Fig. 4). This indicates that the accuracy of shape string prediction affects the performance of γ-turn predictor directly.

Prediction results using different variables

We make many combinations of the available variables to test the joint effect and the prediction results are shown in Table 2.

We can see that the performance of our method is improved by the adding shape string variable. The prediction accuracy of shape string generated by our shape string predictor using SD$_{\_30}$ is 79.8% (Fig. 4) whereas the Q_{total} and MCC can reach 78.6% and 0.23, which improves 25.2% and 0.12 with and without shape string, respectively. It is also evident that the performance improves with more accurate shape string. The Q_{total} and MCC reach 92.2% and 0.38 when using PSSMs, PSS, and shape strings generated using SD$_{\_100}$, respectively. This is the best result obtained by our method. We also notice that the Q_{observed} is lower when adding shape string, whereas the $Q_{\mathrm{predicted}}$ is higher, which demonstrates that many of the predicted γ-turns are true γ-turns.

Table 1 Prediction results using shape string generated from different structure database as input

SSD	Q_{total}	$Q_{\mathrm{predicted}}$	Q_{observed}	MCC
SSD$_{\_30}$	91.5	22.5	51.5	0.30
SSD$_{\_60}$	91.5	22.8	59.2	0.33
SSD$_{\_90}$	92.1	24.7	61.3	0.35
SSD$_{\_100}$	92.2	25.3	63.0	0.37
SSD$_{\_real}$	96.0	44.6	79.0	0.58

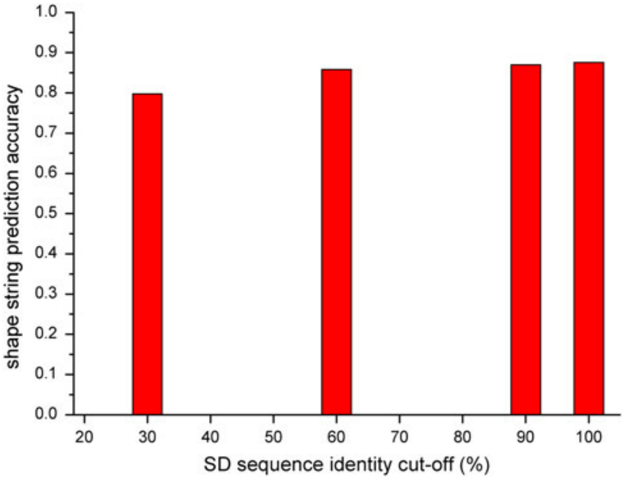

Fig. 4 Prediction accuracy for shape string generated by our predictor searching against structure database for different sequence identity cut-offs

Table 2 Prediction results using different variables as input

Variables	Q_{total}	$Q_{\mathrm{predicted}}$	Q_{observed}	MCC
PSSMs	60.5	5.6	70.0	0.11
PSSMs + PSS	53.4	5.4	79.4	0.11
PSSMs + PSS + SSD$_{\_30}$	78.6	10.9	76.4	0.23
PSSMs + PSS + SSD$_{\_60}$	86.0	15.5	72.9	0.29
PSSMs + PSS + SSD$_{\_90}$	89.1	19.3	71.0	0.33
PSSMs + PSS + SSD$_{\_100}$	92.2	25.4	67.0	0.38
PSSMs + PSS + SSD$_{\_real}$	97.7	69.3	76.3	0.67

Performance comparison with other competing methods

Table 3 shows the comparison between the present method and other popular γ-turn prediction methods. SNNS, based on neural networks, is generally considered the most reliable and accurate γ-turn prediction method. It can be seen that the MCC is appreciably higher for the present method (0.38) than for SNNS (0.17). The MCC is a robust and balanced performance measure that takes into account both the overpredictions and underpredictions. The accuracy of

Table 3 Performance comparison of the present method and other methods

Methods	Q_{total}	$Q_{\text{predicted}}$	Q_{observed}	MCC
Our approach	92.2	25.4	67.0	0.38
Hu X et al.'s SVM	61.0	6.8	91.4	0.18
SNNS	74.0	6.3	83.2	0.17
GTSVM	67.4	6.3	64.7	0.12
WEKA-logistic regression	62.6	5.6	65.1	0.12
WAKE-naïve Bayes	57.4	5.0	65.4	0.11
GOR	75.5	6.1	45.5	0.09
Sequence coupled model	57.8	5.9	43.2	0.08
WAKE-J48 classifier	92.6	5.0	7.2	0.03

The results of sequence couple model, GOR, WAKE, SNNS are obtained from (Kaur and Raghava 2003), the results of GTSVM are obtained from (Pham et al. 2005) and Hu X et al.'s SVM is from (Hu and Ll 2008)

Q_{total} and $Q_{\text{predicted}}$ is 18.2 and 19.1% higher than the SNNS, respectively. Although the Q_{observed} is lower than SNNS and Hu X et al.'s SVM, their $Q_{\text{predicted}}$ are very low. This indicates that many fragments of the predicted γ-turns are false positive.

Discussion

Shape string as new variable was introduced for γ-turn prediction and the excellent performance of the present method demonstrates that it is highly suitable for γ-turn prediction. As described by Hovmöller et al. (2002), the shape string may be useful for protein-folding prediction. We can see that the γ-turn belongs to the turn region in the Ramachandran plots (see Fig. 1 in Ison et al. 2005) and shape string reflects the clustering information, which may be the main reason why shape string plays an important role on the γ-turn prediction.

Our predictor is simple but efficient. It combines the shape strings of homologous sequence fragments matched to the queried sequence as the predicted shape strings, but it uses the available information in the existing database and the structure information at the same time. With the increasing number of proteins in PDB, the shape string produced by our predictor will be closer to the real shape string, and the result will be more accurate approaching the result with addition of the real shape string as input (Table 2). We believe that many other algorithms will be developed to improve the prediction accuracy of shape string, and further develop the γ-turn prediction work in the future.

Meanwhile, we understand shape string is only one kind of the secondary structures. It is verified by many researchers that accurate secondary structure prediction can greatly improve the accuracy of the γ-turn prediction. Protein structure alignment algorithms have already been proposed (Wang et al. 2010) and this method has also been employed in protein secondary structure prediction and achieved much better result (Montgomerie et al. 2006). Both accurate secondary structure and shape string achieved by structure alignment are used as variables, will give a more accurate results of tight turn prediction.

Conclusion

In this paper, we present a novel method of γ-turn prediction. We make two innovations, using G-mean metric as optimal criterion for SVM on the imbalanced γ-turn dataset and introducing the new variable shape string. The obtained results denote that these innovations are suitable. The mainly reason for the good performance of the present method is the shape string, which is predicted based on structure alignment and reflects the clustering information.

Generally, progress in research into prediction of γ-turn comprises several stages. In the first stage, amino acid composition concomitant with different algorithms were proposed and played a main role. Then, PSSMs and predicted secondary structure were subsequently introduced into modeling and more powerful algorithms such as SVM were launched. A new stage is now emerging that involves structure alignment methods. More useful algorithms will be proposed to fill the gap between abundant and accurate protein structure data and the relatively low accuracy of γ-turn prediction.

Acknowledgments The authors would like to thank the National Natural Science Foundation of China (20675057, 20705024) for financial support.

References

Alkorta I, Suarez ML, Herranz R, GonzalezMuniz R, GarciaLopez MT (1996) Similarity study on peptide gamma-turn conformation mimetics. J Mol Model 2:16–25

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Anand A, Pugalenthi G, Fogel GB, Suganthan PN (2010) An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids 39:1385–1391

Barandela R, Sanchez JS, Garcia V, Rangel E (2003) Strategies for learning in class imbalance problems. Pattern Recogn 36:849–851

Bystrov VF, Portnova SL, Tsetlin VI, Ivanov VT, Ovchinnikov YA (1969) Conformational studies of peptide systems. The rotational states of the NH–CH fragment of alanine dipeptides by nuclear magnetic resonance. Tetrahedron 25:493–515

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for the classification and prediction of beta-turn types. J Pept Sci 8:297–301

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

Chou KC (1997a) Prediction and classification of alpha-turn types. Biopolymers 42:837–853

Chou KC (1997b) Prediction of beta-turns. J Pept Res 49:120–144

Chou KC (2000) Prediction of tight turns and their types in proteins. Anal Biochem 286:1–16

Chou KC, Blinn JR (1997) Classification and prediction of beta-turn types. J Protein Chem 16:575–595

Chou KC, Cai YD (2002) Using functional domain composition and support vectormachines for prediction of protein subcellular location. J Biol Chem 277:45765–45769

DiFrancesco V, Garnier J, Munson PJ (1996) Improving protein secondary structure prediction with aligned homologous sequences. Protein Sci 5:106–113

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120

Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J Mol Biol 198:425–443

Guruprasad K, Rajkumar S (2000) Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. J Biosci 25:143–156

Guruprasad K, Shukla S, Adindla S, Guruprasad L (2003) Prediction of gamma-turns from amino acid sequences. J Pept Res 61:243–251

Hovmöller S, Zhou T (2004) Protein shape strings and DNA sequences [Online]. Available: http://www.fos.su.se/∼pdbdna/pdb_shape_dna.html

Hovmöller S, Zhou T, Ohlson T (2002) Conformations of amino acids in proteins. Acta Crystallogr D 58:768–776

Hu XZ, Li QZ (2008) Using support vector machine to predict beta- and gamma-turns in proteins. J Comput Chem 29:1867–1875

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

Ison RE, Hovmöller S, Kretsinger RH (2005) Proteins and their shape strings. An exemplary computer representation of protein structure. IEEE Eng Med Biol Mag 24:41–49

Jahandideh S, Sarvestani AS, Abdolmaleki P, Jahandideh M, Barfeie M (2007) gamma-turn types prediction in proteins using the support vector machines. J Theor Biol 249:785–790

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Kaur H, Raghava GPS (2002) An evaluation of {beta}-turn prediction methods. Bioinformatics 18:1508–1514

Kaur H, Raghava GPS (2003) A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 12:923–929

Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. Morgan Kaufmann 179–186. doi:10.1.1.43.4487

Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinform 7:301

Pham TH, Satou K, Ho TB (2005) Support vector machines for prediction and analysis of beta and gamma-turns in proteins. J Bioinform Comput Biol 3:343–358

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339

Robert MK, Holte R, Matwin S (1997) Learning when negative examples abound. Springer, Berlin, pp 146–153. doi: 10.1.1.36.88

Rose GD, Gierasch LM, Smith JA (1985) Turns in peptides and proteins. Adv Protein Chem 37:1–109

Shepherd AJ, Gorse D, Thornton JM (1999) Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci 8:1045–1055

Wang L, Wu LY, Wang Y, Zhang XS, Chen LN (2010) SANA: an algorithm for sequential and non-sequential protein structure alignment. Amino Acids 39:417–425

Wrtten IH, Frank E (1999) Data mining: practical machine learning tools and techniques with java implementations. Morgan Kaufmann, San Francisco

Wu G, Chang EY (2003) Class-boundary alignment for imbalanced dataset learning. doi:10.1.1.94.9007

Zell A, Mamier G (1997) Neural Network Simulator, Version 4.2. University of Stuttgart, Stuttgart

Zhang Q, Yoon S, Welsh WJ (2005) Improved method for predicting beta-turn using support vector machine. Bioinformatics 21:2370–2374

Zhou TP, Shu NJ, Hövmoller S (2010) A novel method for accurate one-dimensional protein structure prediction based on fragment matching. Bioinformatics 26:470–477