Amino Acids (2015) 47:461–468
DOI 10.1007/s00726-014-1878-9

ORIGINAL ARTICLE

Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination

Peiying Tao · Taigang Liu · Xiaowei Li · Lanming Chen

Received: 8 July 2014 / Accepted: 17 November 2014 / Published online: 13 January 2015

Abstract Knowledge of structural class plays an important role in understanding protein folding patterns. As a transitional stage in recognition of three-dimensional structure of a protein, protein structural class prediction is considered to be an important and challenging task. In this study, we firstly introduce a feature extraction technique which is based on tri-grams computed directly from position-specific scoring matrix (PSSM). A total of 8,000 features are extracted to represent a protein. Then, support vector machine-recursive feature elimination (SVM-RFE) is applied for feature selection and reduced features are input to a support vector machine (SVM) classifier to predict structural class of a given protein. To examine the effectiveness of our method, jackknife tests are performed on six widely used benchmark datasets, i.e., Z277, Z498, 1189, 25PDB, D640, and D1185. The overall accuracies of 97.1, 98.6, 92.5, 93.5, 94.2, and 95.9 % are achieved on these datasets, respectively. Comparison of the proposed method with other prediction methods shows that our method is very promising to perform the prediction of protein structural class.

Keywords Tri-gram · Position-specific scoring matrix · Protein structural class · Support vector machine · Feature selection

P. Tao · X. Li · L. Chen (✉)
College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China
e-mail: lmchen@shou.edu.cn

T. Liu (✉)
College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
e-mail: tgliu@shou.edu.cn

Introduction

The structural class knowledge of a given protein can provide a useful bit of information about its overall structure, which has played an extremely important role in understanding protein function. Since Levitt and Chothia (1976) proposed the concept of protein structural class, known protein structures are classified into four categories: all-α, all-β, α/β, and $\alpha + \beta$. Experimental approaches to determine the structure information of a protein, including X-ray Diffraction and Nuclear Magnetic Resonance, are costly and time-consuming, and thus are not capable of completely meeting researchers' demands (Li et al. 2014). Therefore, fast computational approaches are brought to tackle this issue.

During the past three decades, a variety of methods have been proposed to predict the structural class of protein. These methods generally comprise two steps: (1) protein sequence feature extraction; (2) the selection of classification algorithm. For the former step, many sequence features have been used to represent protein sequences, including amino acid composition (AAC) (Nakashima et al. 1986; Chou 1999), pseudo amino acid composition (PseAA) (Chou 2001; Lin and Li 2007), polypeptide composition (Luo et al. 2002), functional domain composition (Chou and Cai 2004), amino acid sequence reverse encoding (Deschavanne and Tuffery 2008; Mizianty and Kurgan 2009), PSI-BLAST profile (Chen et al. 2008; Liu et al. 2010, 2012), and predicted secondary structure information (Kurgan et al. 2008b; Yang et al. 2010). Recently, several integrated feature extraction methods have been developed to further improve the prediction performance (Dehzangi et al. 2014; Li et al. 2014). For instance, by fusing multi-source information from PSI-BLAST profile, PROFEAT,

and Gene Ontology annotation, PSSP-RFE method (Li et al. 2014) shows significant enhancements in prediction accuracies compared with other existing methods especially for low-similarity datasets. For the latter step, several machine learning algorithms have been used to accomplish the protein structural class prediction, such as neural network (Cai and Zhou 2000), support vector machine (SVM) (Cai and Zhou 2000; Cai et al. 2001, 2002; Chen et al. 2006a; Anand et al. 2008; Li et al. 2008; Liu et al. 2010; Yang et al. 2010), fuzzy clustering (Shen et al. 2005), Bayesian-based learners (Wang and Yuan 2000), logistic regression (Kurgan and Homaeian 2006; Kedarisetti et al. 2006b), rough sets (Cao et al. 2006), and information discrepancy (Jin et al. 2003; Kedarisetti et al. 2006b; Kurgan and Homaeian 2006). Besides, some complex classification models have also been applied, such as ensembles (Kedarisetti et al. 2006a), bagging (Dong et al. 2006), and boosting (Feng et al. 2005). Among these classification algorithms, SVM is the most widespread application and demonstrates quite promising results (Chen et al. 2006a; Kurgan et al. 2008b; Liu et al. 2012).

In the literature, Paliwal et al. (2014) have used tri-gram features for protein fold recognition. They have computed the tri-gram features directly from position-specific scoring matrix (PSSM) generated by PSI-BLAST program (Altschul et al. 1997). Since there are 20 amino acids of interest, there will be $20 \times 20 \times 20 = 8{,}000$ combinations of tri-grams, giving an 8,000 dimensional feature vector for a given protein sequence. However, the dimensionality of the tri-gram feature vector is quite large, which may include some irrelevant and redundant features. Thus, computing all tri-gram features from PSSM as the feature vector is not an effective way of capturing the information.

In the present work, we aim to combine tri-gram features and a feature selection method to improve the prediction accuracy of protein structural class. At first, we use tri-gram features of PSSM to represent each protein sequence. Then, instead of computing all tri-gram features from PSSM (Paliwal et al. 2014), we do a feature selection step with support vector machine-recursive feature elimination (SVM-RFE). Feature selection can identify and remove as much irrelevant and redundant information as possible. Therefore, reduced features are input to an SVM to perform the prediction, and more useful information would be retrieved. According to the results of jackknife cross-validation tests on six widely used benchmark datasets, the current method presents satisfying prediction accuracies in comparison with existing methods. The datasets used in this study and the source code for implementing the algorithm are freely available to the academic community at http://xxxy.shou.edu.cn/bioinform/Trigram-PSSM-RFE/index.html.

Materials and methods

Protein datasets

To explore the impact of sequence similarity on the performance of the current method, we adopt two groups of datasets. One group is the high-similarity datasets, namely Z277 and Z498, which were originally generated by Zhou (1998) and were used in many previous studies (Zhou 1998; Chou 2005; Cao et al. 2006; Liu et al. 2012). They contain 277 and 498 protein domains, respectively. Another group comprises four low-similarity datasets, 1189 (Wang and Yuan 2000), 25PDB (Kurgan and Homaeian 2006), D640 (Chen et al. 2008), and D1185 (Xia et al. 2012). In these four datasets, pair-wise sequence identities are less than 40, 25, 25, and 40 %, respectively. The detailed information about the six datasets is listed in Table 1.

Tri-gram feature extraction

In this section, we extract tri-gram features from PSSM to represent protein sequences. The extracted features combine neighborhood information of amino acids and evolutionary information from PSSM, which are quite effective for protein structural class prediction. The PSSM is a log-odd matrix of size $L \times 20$, where L is the length of the primary sequence. The element at ith row and jth column is denoted by $p_{i,j}$ which indicates the relative probability of jth amino acid at the ith location of the protein sequence during biological evolution processes. The PSSM elements are mapped to the range of (0, 1) by the following standard sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}},$$

where x is the original PSSM value.

The probability of amino acids triplet which consists of uth, vth, and wth amino acids in order is calculated as follows:

$$T_{u,v,w} = \sum_{i=1}^{L-2} p_{i,u} p_{i+1,v} p_{i+2,w},$$

Table 1 The compositions of six datasets adopted in this study

Dataset	Number of proteins				
	All-α	All-β	α/β	$\alpha + \beta$	Overall
Z277	70	61	81	65	277
Z498	107	126	136	129	498
1189	223	294	334	241	1092
25PDB	443	443	346	441	1673
D640	138	154	177	171	640
D1185	251	258	199	477	1185

where $1 \le u \le 20$, $1 \le v \le 20$, and $1 \le w \le 20$.

This equation generates 8,000 frequencies of occurrences $T_{u,v,w}$. We define the matrix T as the tri-gram occurrence matrix and 8,000 elements can be formulated as follows:

$$F = (T_{1,1,1}, T_{1,1,2}, \ldots, T_{1,1,20}, T_{1,2,1}, \ldots, T_{1,20,20}, T_{2,1,1}, \ldots, T_{2,1,20}, \ldots, T_{2,20,20}, \ldots, T_{20,20,20})',$$

where $'$ is a transpose of the descriptor vector.

These 8,000 elements of three-dimensional matrix $[T_{u,v,w}]$ which correspond to the probabilities of PSSM-based tri-grams are employed to represent the given protein sequence. However, there may have some uncorrelated and superfluous information among the extracted features, which can affect the speed and performance of prediction. Hence, a feature selection method is necessary.

Feature selection by SVM-RFE

Feature selection is the key preprocessing step dealing with dimensionality reduction for pattern recognition. In classification, it is used to find an optimal subset of considered features so that the overall accuracy is increased while the data size is reduced. Since the proposed protein sequence representation contains a huge number of features, SVM-RFE is utilized to reduce the dimensionality and computation complexity in this study, which was initially proposed for cancer classification (Guyon et al. 2002). Firstly, all the feature vectors of proteins for each dataset would be used to construct a feature matrix, where each row represents a sample and each column represents a feature. Then, SVM-RFE algorithm is implemented by training an SVM with a linear kernel to get a ranked list of all features. Finally, top K features with the most relevant ranks are selected to represent a protein.

Support vector machine

In this study, we employ SVM (Vapnik 1995) as a classifier. SVM is considered to be the state-of-the-art machine learning and pattern classification algorithm. Compared to other machine learning methods, SVM has the advantages of high performance, absence of local minima, and the speed ability to deal with multidimensional datasets with complex relationships among the data elements. Here, the publicly available LIBSVM software (Chang and Lin 2011) is used to perform the SVM classifier. Although LIBSVM provides a choice of in-built kernels, such as linear, polynomial, and radial basis function, linear kernel is employed in our experiments because of its simple operation and good performance without parameters optimization.

Performance measures

In statistical prediction, independent dataset test, sub-sampling test, and jackknife test are the most common methods for evaluating the prediction performance. Of these three methods, the jackknife test is considered the most rigorous and objective (Chou and Zhang 1995). Hence, we adopt the jackknife test to validate the predictor. In the jackknife test, each of protein in the dataset is in turn singled out as a tested protein, and the predictor is trained by the remaining proteins. In the present study, three standard performance measures are used to evaluate the prediction accuracy, i.e., accuracy, overall accuracy, and Matthew's Correlation Coefficient (MCC) (Matthews 1975). They are defined by the following formulas:

$$\text{Accuracy}_j = \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j} = \frac{\text{TP}_j}{|C_j|}$$

$$\text{Overall accuracy} = \frac{\sum_j \text{TP}_j}{\sum_j |C_j|}$$

$$\text{MCC}_j = \frac{\text{TP}_j \times \text{TN}_j - \text{FP}_j \times \text{FN}_j}{\sqrt{(\text{TP}_j + \text{FP}_j)(\text{TP}_j + \text{FN}_j)(\text{TN}_j + \text{FP}_j)(\text{TN}_j + \text{FN}_j)}},$$

where TP_j, TN_j, FP_j, FN_j, and $|C_j|$ are the number of true positives, true negatives, false positives, false negatives, and proteins in the structural class C_j, respectively.

Results and discussion

Effect of top K features

In this study, we apply SVM-RFE to get a rank list of 8000 tri-gram features according to their contributions to protein structural class prediction and compute overall accuracies for top K features using tenfold cross-validation tests, where K = 10, 20, 30, …, 600. The results are shown in Fig. 1. From Fig. 1, we find that overall accuracies at top 80 features are the highest for both Z277 and Z498 datasets. The total accuracy of the 25PDB dataset achieves a maximum value when K increases to 500. In addition, favorable accuracies are also obtained at this point for the 1189, D640, and D1185 datasets. To make the proposed descriptor become a uniform representation, a 500-dimensional feature set is constructed and used to perform the protein structural class prediction in the rest of this study.

Prediction accuracies of our method on six working datasets

Six widely used datasets are adopted to estimate the performance of our method by the jackknife test, including

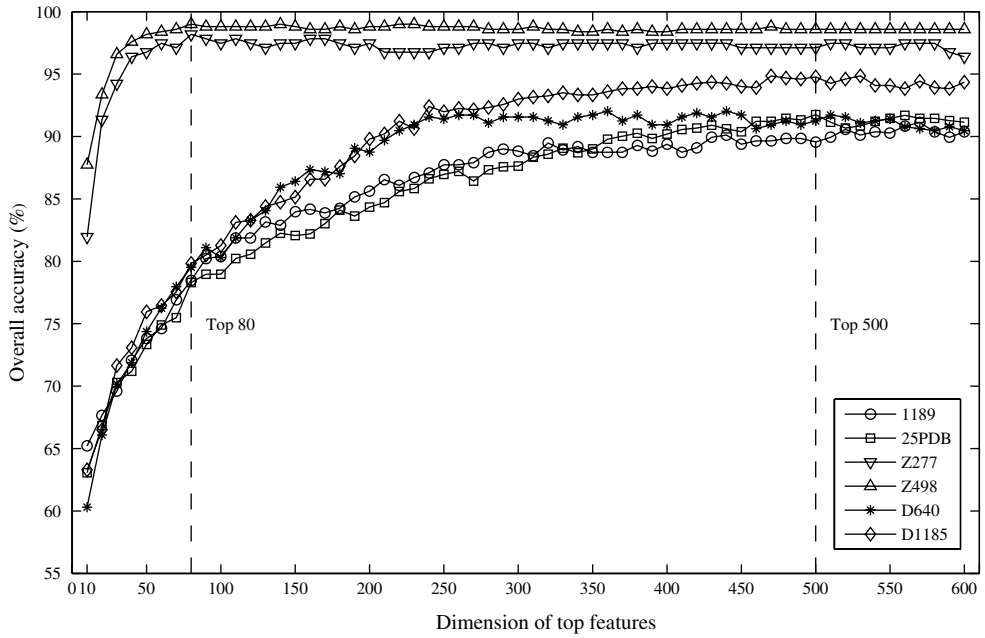

Fig. 1 This graph shows how different top K features affect the overall accuracies

Table 2 Prediction performances on six datasets by our method

Dataset	Accuracy (%)					MCC			
	All-α	All-β	α/β	$\alpha+\beta$	Overall	All-α	All-β	α/β	$\alpha+\beta$
Z277	98.6	96.7	96.3	96.9	97.1	0.97	0.95	0.95	0.98
Z498	98.1	100	97.1	99.2	98.6	0.98	0.98	0.97	0.99
1189	94.6	94.2	92.5	88.4	92.5	0.89	0.91	0.91	0.88
25PDB	97.1	93.9	89.9	92.3	93.5	0.93	0.92	0.90	0.90
D640	93.5	97.4	93.2	93.0	94.2	0.90	0.95	0.91	0.93
D1185	96.0	96.1	92.0	97.3	95.9	0.94	0.96	0.91	0.94

high-similarity datasets Z277 and Z498, and low-similarity datasets 1189, 25PDB, D640, and D1185. In this section, we select the accuracy, MCC of each class and overall accuracy as performance measures that are summarized in Table 2. The overall accuracies of our method are 97.1, 98.6, 92.5, 93.5, 94.2, and 95.9 % for these six datasets, respectively. It is observed that the prediction performances of high-similarity datasets are better (with accuracies higher than 96 %, and MCC values not lower than 0.95 for all classes) than those of low-similarity datasets. This implies that protein sequence similarity within the training and testing datasets has a significant impact on the prediction performance of protein structural class. In addition, for low-similarity datasets, the accuracies for predicting α/β and $\alpha + \beta$ classes are relatively low when compared with the other classes. The reason of low prediction accuracy may be its non-negligible overlap with the other classes.

Comparison with other prediction methods

To demonstrate the effectiveness of the proposed method, we compare it with some previous methods based on the same datasets. The results of other methods are obtained directly from the original papers (Tables 3, 4, 5, 6, 7, 8).

For two high-similarity datasets, Z277 and Z498, our method reaches the overall accuracies of 97.1 and 98.6 % (Tables 3, 4), which are higher than those of the other methods except for the PSSP-RFE method (Li et al. 2014). It should be pointed out that PSSP-RFE additionally extracts physicochemical features and Gene Ontology annotation features as its input and may be unable to predict the structural class for a few proteins due to a lack of their Gene Ontology numbers. However, our method also obtains satisfactory prediction accuracies when only tri-gram features extracted from PSSM are employed. In addition, it is worth noted that the other methods based on

Table 3 Comparison of different methods by the jackknife test for the Z277 dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
Rough sets (Cao et al. 2006)	77.1	77.0	93.8	66.2	79.4
Information-theoretical approach (Zheng et al. 2010)	87.1	80.3	93.8	67.7	83.0
LogitBoost (Feng et al. 2005)	81.4	88.5	92.6	72.3	84.1
IGA-SVM (Li et al. 2008)	84.3	88.5	92.6	70.7	84.5
CWT-PCA-SVM (Li et al. 2009)	85.7	90.2	87.7	80.1	85.9
IB1 (Chen et al. 2008)	89.7	88.1	92.2	80.0	87.7
SVM fusion (Chen et al. 2006b)	85.7	90.2	93.8	80.0	87.7
AAC-PSSM-AC (Liu et al. 2012)	88.6	95.1	97.5	81.5	91.0
PSSP-RFE (Li et al. 2014)	100	98.3	100	100	99.6
Our method	98.6	96.7	96.3	96.9	97.1

Table 4 Comparison of different methods by the jackknife test for the Z498 dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
Rough sets (Cao et al. 2006)	87.9	91.3	97.1	86.0	90.8
SVM fusion (Chen et al. 2006b)	99.1	96.0	80.9	91.5	91.4
Information-theoretical approach (Zheng et al. 2010)	95.3	93.7	97.8	88.3	93.8
IGA-SVM (Li et al. 2008)	96.3	93.6	97.8	89.2	94.2
LogitBoost (Feng et al. 2005)	92.6	96.0	97.1	93.0	94.8
CWT-PCA-SVM (Li et al. 2009)	94.4	96.8	97.0	92.3	95.2
IB1 (Chen et al. 2008)	95.0	95.8	97.8	94.2	95.7
AAC-PSSM-AC (Liu et al. 2012)	94.4	96.8	97.8	93.8	95.8
PSSP-RFE (Li et al. 2014)	100	99.2	100	100	99.8
Our method	98.1	100	97.1	99.2	98.6

Table 5 Performance comparison of different methods on the 1189 dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
AADP-PSSM (Liu et al. 2010)	69.1	83.7	85.6	35.7	70.7
AAC-PSSM-AC (Liu et al. 2012)	80.7	86.4	81.4	45.2	74.6
Comb_11,10,6[a] (Dehzangi et al. 2013)	80.2	83.6	85.4	44.6	74.8
SCPRED (Kurgan et al. 2008a)	89.1	86.7	89.6	53.8	80.6
LCC-PSSM (Ding et al. 2014)	89.2	88.8	85.6	58.5	81.2
RKS-PPSC (Yang et al. 2010)	89.2	86.7	82.6	65.6	81.3
MODAS (Mizianty and Kurgan 2009)	92.3	87.1	87.9	65.4	83.5
PSSM-SPINE-S (Dehzangi et al. 2014)	98.2	91.5	83.8	72.2	86.3
PSSP-RFE (Li et al. 2014)	94.9	96.8	96.6	97.1	96.4
Our method	94.6	94.2	92.5	88.4	92.5

[a] The result is evaluated using tenfold cross-validation test

PSSM features, IB1 (Chen et al. 2008), and AAC-PSSM-AC (Liu et al. 2012), also get the high prediction accuracies. This indicates that PSI-BLAST profile indeed contains quite important information for protein structural class prediction.

For low-similarity datasets, our method also attains better prediction results compared to most of the existing methods. For the 1189 dataset (Table 5), only two methods provide the overall accuracies over 90 %. One is our method, and the other is PSSP-RFE method (Li et al. 2014). In addition, PSSM-SPINE-S (Dehzangi et al. 2014) method achieves the third best results. It is noteworthy that PSSM-SPINE-S method combines PSSM features with predicted secondary structure features to improve the performance. Comparison with PSSM-SPINE-S method reveals that predicted secondary structure information provides an important complementary role for predicting protein structural class. As shown in Table 6, results on the 25PDB dataset are mainly consistent with those on the 1189 dataset. Our method, together with the PSSP-RFE method, performs better than the other

Table 6 Performance comparison of different methods on the 25PDB dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
AADP-PSSM (Liu et al. 2010)	83.3	78.1	76.3	54.4	72.9
Comb_11,10,6[a] (Dehzangi et al. 2013)	86.1	80.8	80.6	60.1	76.7
AAC-PSSM-AC (Liu et al. 2012)	85.3	81.7	73.7	55.3	74.1
LCC-PSSM (Ding et al. 2014)	91.7	80.8	79.8	64.0	79.0
SCPRED (Kurgan et al. 2008a)	92.6	80.1	74.0	71.0	79.7
MODAS (Mizianty and Kurgan 2009)	92.3	83.7	81.2	68.3	81.4
RKS-PPSC (Yang et al. 2010)	92.8	83.3	85.8	70.1	82.9
PSSM-SPINE-S (Dehzangi et al. 2014)	96.8	93.7	90.1	87.0	92.2
PSSP-RFE (Li et al. 2014)	94.9	95.5	95.8	91.4	94.3
Our method	97.1	93.9	89.9	92.3	93.5

[a] The result is evaluated using tenfold cross-validation test

Table 7 Performance comparison of different methods on the D640 dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
SCEC (Chen et al. 2008)	73.9	61.0	81.9	33.9	62.3
LCC-PSSM (Ding et al. 2014)	92.8	88.3	85.9	66.1	82.7
RKS-PPSC (Yang et al. 2010)	89.1	85.1	88.1	71.4	83.1
OET-KNN (Hayat et al. 2014)	86.4	85.3	90.4	90.6	88.4
PseAA-SSP (Wang et al. 2014)	95.7	89.6	89.3	90.1	90.9
PSSP-RFE (Li et al. 2014)	95.5	96.6	97.0	93.9	95.7
Our method	93.5	97.4	93.2	93.0	94.2

Table 8 Performance comparison of different methods on the D1185 dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha+\beta$	Overall
MLR model (Xia et al. 2012)	95.6	81.0	78.9	71.9	80.1
PSSP-RFE (Li et al. 2014)	83.1	82.7	79.8	88.5	84.6
Our method	96.0	96.1	92.0	97.3	95.9

methods, with an overall accuracy of 93.5 %. In detail, our method shows the best performances for the all-α and $\alpha+\beta$ classes in comparison with the other methods. The accuracies of the all-β and α/β classes are slightly lower than those of PSSM-SPINE-S and PSSP-RFE, but still higher than those of the other methods. Besides, good performances are also obtained on the other two low-similarity datasets D640 and D1185 (Tables 7, 8). Especially for the D1185 dataset, our method shows greater improvements compared with MLR model (Xia et al. 2012) and PSSP-RFE method. In summary, the above comparison results suggest that our method is very promising and could provide a cost-alternative to predict protein structural class. Moreover, the prediction performance may be further improved by integrating features from physicochemical property, Gene Ontology, and predicted secondary structure in the future.

Conclusions

As a key step in recognition of three-dimensional structure of a protein, protein structural class prediction becomes an important and challenging task. In this study, we have introduced a tri-gram feature extraction technique and a recursive feature selection scheme based on linear kernel SVM in order to select the optimal features for predicting protein structural class. A total of 8,000 features are extracted to represent a protein by computing tri-gram probabilities of PSSM, which integrate neighborhood information and evolutionary information. SVM-RFE is then used to rank these feature vectors and select optimal features. Finally, SVM classifier is applied for predicting protein structural class based on selected features. Jackknife cross-validation tests on six working datasets show that our method is superior or comparable to other existing methods in identifying protein structural class. Comparison results suggest that the proposed method could serve as a very useful tool for prediction of protein structural class. A web server which provides the source code for implementing the algorithm and the datasets used in our work is freely available at http://xxxy.shou.edu.cn/bioinform/Trigram-PSSM-RFE/index.html. In addition, we shall make efforts in our future work to provide the online prediction service.

Acknowledgments This work was supported by the Innovation Program of Shanghai Municipal Education Commission (No. 13YZ098), the Foundation for University Youth Teachers of Shanghai (No. ZZhy12028), the National Natural Science Foundation of China (Nos. 31271830, 41376135), and the Doctoral Fund of Shanghai Ocean University.

Conflict of interest The authors declare no conflict of interest related to this study.

References

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. Nucleic Acids Res 25(17):3389–3402. doi:10.1093/nar/25.17.3389

Anand A, Pugalenthi G, Suganthan PN (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J Theor Biol 253(2):375–380. doi:10.1016/j.jtbi.2008.02.031

Cai YD, Zhou GP (2000) Prediction of protein structural classes by neural network. Biochimie 82(8):783–785
Cai YD, Liu XJ, Xu X, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC Bioinform 2:3. doi:10.1186/1471-2105-2-3
Cai YD, Liu XJ, Xu XB, Chou KC (2002) Prediction of protein structural classes by support vector machines. Comput Chem 26(3):293–296. doi:10.1016/s0097-8485(01)00113-9
Cao YF, Liu S, Zhang LD, Qin J, Wang J, Tang KX (2006) Prediction of protein structural class with Rough Sets. BMC Bioinform 7:20. doi:10.1186/1471-2105-7-20
Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2(3):27. doi:10.1145/1961189.1961199
Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243(3):444–448. doi:10.1016/j.jtbi.2006.06.025
Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357(1):116–121. doi:10.1016/j.ab.2006.07.022
Chen K, Kurgan LA, Ruan JS (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem 29(10):1596–1604. doi:10.1002/Jcc.20918
Chou KC (1999) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264(1):216–224. doi:10.1006/bbrc.1999.1325
Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43(3):246–255. doi:10.1002/prot.1035
Chou KC (2005) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6(5):423–436. doi:10.2174/138920305774329368
Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321(4):1007–1009. doi:10.1016/j.bbrc.2004.07.059
Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30(4):275–349. doi:10.3109/10409239509083488
Dehzangi A, Paliwal K, Sharma A, Dehzangi O, Sattar A (2013) A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. IEEE/ACM Trans Comput Biol Bioinform 10(3):564–575. doi:10.1109/TCBB.2013.65
Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A (2014) Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC Genomics 15(Suppl 1):S2. doi:10.1186/1471-2164-15-s1-s2
Deschavanne P, Tuffery P (2008) Exploring an alignment free approach for protein classification and structural class prediction. Biochimie 90(4):615–625. doi:10.1016/j.biochi.2007.11.004
Ding S, Yan S, Qi S, Li Y, Yao Y (2014) A protein structural classes prediction method based on PSI-BLAST profile. J Theor Biol 353:19–23. doi:10.1016/j.jtbi.2014.02.034
Dong L, Yuan Y, Cai Y (2006) Using bagging classifier to predict protein domain structural class. J Biomol Struct Dyn 24(3):239–242
Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334(1):213–217. doi:10.1016/j.bbrc.2005.06.075
Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422. doi:10.1023/A:1012487302797
Hayat M, Tahir M, Khan SA (2014) Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J Theor Biol 346(7):8–15. doi:10.1016/j.jtbi.2013.12.015
Jin LX, Fang WW, Tang HW (2003) Prediction of protein structural classes by a new measure of information discrepancy. Comput Biol Chem 27(3):373–380. doi:10.1016/S1476-9271(02)00087-7
Kedarisetti KD, Kurgan L, Dick S (2006a) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348(3):981–988. doi:10.1016/j.bbrc.2006.07.141
Kedarisetti KD, Kurgan L, Dick S (2006b) A comment on—"Prediction of protein structural classes by a new measure of information discrepancy". Comput Biol Chem 30(5):393–394. doi:10.1016/j.compbiolchem.2006.06.003
Kurgan LA, Homaeian L (2006) Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn 39(12):2323–2343. doi:10.1016/j.patcog.2006.02.014
Kurgan L, Cios K, Chen K (2008a) SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinform 9:226. doi:10.1186/1471-2105-9-226
Kurgan LA, Zhang T, Zhang H, Shen SY, Ruan JS (2008b) Secondary structure-based assignment of the protein structural classes. Amino Acids 35(3):551–564. doi:10.1007/s00726-008-0080-3
Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261(5561):552–558. doi:10.1038/261552a0
Li ZC, Zhou XB, Lin YR, Zou XY (2008) Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35(3):581–590. doi:10.1007/s00726-008-0084-z
Li ZC, Zhou XB, Dai Z, Zou XY (2009) Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37(2):415–425. doi:10.1007/s00726-008-0170-2
Li L, Cui X, Yu S, Zhang Y, Luo Z, Yang H, Zhou Y, Zheng X (2014) PSSP-RFE: Accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations. PLoS One 9(3):e92863. doi:10.1371/journal.pone.0092863
Lin H, Li QZ (2007) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28(9):1463–1466. doi:10.1002/Jcc.20554
Liu TG, Zheng XQ, Wang J (2010) Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. Biochimie 92(10):1330–1334. doi:10.1016/j.biochi.2010.06.013
Liu T, Geng X, Zheng X, Li R, Wang J (2012) Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42(6):2243–2249. doi:10.1007/s00726-011-0964-5
Luo RY, Feng ZP, Liu JK (2002) Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 269(17):4219–4225. doi:10.1046/j.1432-1033.2002.03115.x
Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405(2):442–451. doi:10.1016/0005-2795(75)90109-9
Mizianty MJ, Kurgan L (2009) Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinform 10:24. doi:10.1186/1471-2105-10-414
Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99(1):153–162
Paliwal KK, Sharma A, Lyons J, Dehzangi A (2014) A tri-gram based feature extraction technique using linear probabilities of position

specific scoring matrix for protein fold recognition. IEEE Trans Nanobiosci 13(1):44–50. doi:10.1109/tnb.2013.2296050

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334(2):577–581. doi:10.1016/j.bbrc.2005.06.128

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

Wang ZX, Yuan Z (2000) How good is prediction of protein structural class by the component-coupled method? Proteins 38(2):165–175. doi:10.1002/(sici)1097-0134(20000201)38:2<165::aid-prot5>3.0.co;2-v

Wang J, Li Y, Liu X, Dai Q, Yao Y, He P (2014) High-accuracy prediction of protein structural classes using PseAA structural properties and secondary structural patterns. Biochimie 101:104–112. doi:10.1016/j.biochi.2013.12.021

Xia X-Y, Ge M, Wang Z-X, Pan X-M (2012) Accurate prediction of protein structural class. PLoS One 7(6):e37653. doi:10.1371/journal.pone.0037653

Yang JY, Peng ZL, Chen X (2010) Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinform 11(Suppl 1):10. doi:10.1186/1471-2105-11-s1-s9

Zheng X, Li C, Wang J (2010) An information-theoretic approach to the prediction of protein structural class. J Comput Chem 31(6):1201–1206. doi:10.1002/jcc.21406

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17(8):729–738. doi:10.1023/a:1020713915365