MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 765-784

ISSN 0340 - 6253

Prediction of Protein Structural Class Based on Different Autocorrelation Descriptors of Position–Specific Scoring Matrix

Yunyun Liang*, Sanyang Liu, Shengli Zhang

School of Mathematics and Statistics, Xidian University, Xi'an 710071, P. R. China

(Received December 14, 2014)

Abstract

Prediction of protein structural class for low-similarity sequences remains a complicated and challenging task in the current bioinformatics. Features extracted based solely on the position-specific scoring matrix (PSSM) have played a significant role in improving the prediction accuracy. In this study, we propose a novel model called MBMGAC-PSSM by fusing PSSM and three autocorrelation descriptors: normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation. Then a 560-dimensional feature vector is constructed. Finally, 175 features are selected using principal component analysis (PCA) on the 1189 dataset. Rigorous jackknife cross-validation tests are performed on three widely used low-similarity benchmark datasets: 1189, 25PDB and 640. Our proposed model achieves the competitive performance on prediction accuracies and also outperforms the other existing PSSM-based methods. The fact shows that our approach can be used as a potential candidate for the accurate prediction of protein structural class.

1 Introduction

Knowledge of protein structural class can provide useful information to understand protein folding patterns[1], and play a central role in improving the prediction quality of protein secondary structure contents, protein tertiary structure, and protein function[2–6]. According to the concept of protein structural class originally introduced by Levitt and Chothia[7], proteins can be categorized into four major structural classes: all-α, all-β,

*Corresponding author. Tel./Fax:+86-29-88202860. E-mail: yunyunliang88@163.com

α/β, and $\alpha+\beta$. The all-α and all-β proteins are mainly formed by helices and strands, respectively. The α/β protein mixes both helices and mostly parallel strands, and the $\alpha+\beta$ protein mixes both helices and mostly antiparallel strands.

During the last two decades, various important efforts that have been made to develop a powerful computational system to tackle this problem. Prediction of protein structural class is a typical and traditional pattern recognition problem, which is generally performed in three main steps: feature extraction, feature selection and model selection for classification. Among the three steps, feature extraction is the most critical and challenging step for the successful improvement of protein structural class prediction. Models widely used include amino acid composition (AAC)[8–11], pseudo-amino acid composition (PseAAC)[12–14], polypeptide composition[15, 16], functional domain composition [17], PSI-BLAST profile[18, 19] and predicted protein secondary structure[20, 21]. In order to decease computation complexity and pick out the actual informative features, a feature selection step is necessary. Widely used feature selection algorithms by researchers include principal component analysis (PCA)[22], SVM-RFE[23], wrapper and filter[24] and so on. Finally, many advanced classification algorithms have been used to implement the protein structural class prediction, such as neural network[25], support vector machine (SVM)[26, 27], fuzzy clustering[28], Bayesian classification[29] and rough sets[30].

Recently, the protein structural class prediction problem especially for low homologous protein sequences, has attracted more attention and its prediction accuracy has been increasingly improved. AADP-PSSM[18] method extends the traditional dipeptide composition to PSSM. AAC-PSSM-AC[19] combines auto covariance and PSSM to extract the evolutionary information. AATP model[31] fuses AAC and transition probability composition from PSSM. In PSSS-PSSM[32], the predicted secondary structure information is employed to perform the prediction with evolutionary information. In MEDP[33], evolutionary difference formula is proposed based on PSSM. The feature extraction methods relying on the the position-specific scoring matrix (PSSM) have played an important role to address this hot issue. However, the information hidden in the PSSM has not been adequately explored, feature extraction remains limited and needs further be improved.

In this study, three widely used autocorrelation descriptors are selected: normalized Moreau-Broto autocorrelation descriptors, Moran autocorrelation descriptors and Geary autocorrelation descriptors[34, 35]. They are all defined based on the scores distribution

of the evolutionary information represented in the form of position-specific scoring matrix along the amino acid sequence. We propose a new comprehensive model called MBMGAC-PSSM by integrating PSSM and three autocorrelation descriptors, which contains not only the evolutionary information, but also the sequence-order information. Meanwhile, a 560-dimensional feature vector is constructed. In order to reduce the influence of noise, we use the principle component analysis (PCA) for feature selection. The 175 dominant features are selected for SVM classifier, which retain most of the information in the sense of maximum variance of the features and minimum reconstruction error. To evaluate our model, jackknife cross-validation test is employed on three widely benchmark datasets, the experimental results show that our model achieves the competitive performance compared with the other evolutionary information-based methods, particularly for low-similarity amino acid sequences.

As demonstrated by a series of recent publications[36–48] in response to the call[51], to establish a really useful sequence-based statistical predictor for a biological system, we need to follow the five guidelines: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) develop a powerful algorithm to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

2 Materials and methods

2.1 Datasets

In order to facilitate the comparison with the previous works, three popular benchmark datasets are used to evaluate the performance of our method: the 1189 dataset[29], the $25PDB$ dataset[49] and the 640 dataset[50], with sequence similarity lower than 40% , 25% and 25%, respectively. The 1189 dataset contains 1092 protein domains, consisting of 223 all-α class proteins, 294 all-β class proteins, 334 α/β class proteins and 241 $\alpha+\beta$ class proteins. The $25PDB$ dataset includes 1673 protein domains, of which 443 is all-α class proteins, 443 is all-β class proteins, 346 is α/β class proteins and 441 is $\alpha+\beta$ class proteins. Referring to the 640 dataset, which contains 640 protein domains, consisting of

138 all-α class proteins, 154 all-β class proteins, 177 α/β class proteins and 171 $\alpha+\beta$ class proteins.

2.2 Feature extraction

To develop a powerful predictor for the protein structural class, the key is how to effectively define a feature vector to formulate the statistical samples concerned. According to Eq.(6) of Chou(2011) [51], the feature vector for any protein, peptide or biological sequence is just the general form of pseudo-amino acid composition or PseAAC[52] that can be formulated as

$$P = (\psi_1, \psi_2, \cdots, \psi_\mu, \cdots, \psi_\Omega)^T \tag{2.1}$$

where T is the transpose operator, while the subscript Ω is an integer and its value as well as the components $\psi_1, \psi_2, \cdots$ will depend on how to extract the desired information from the amino acid sequence of P. In this study, we use the various features extracted from the evolutionary information-based methods, and Ω=560.

2.2.1 Position–specific scoring matrix

To represent a protein sample P with L amino acid residues by its evolution information, position-specific scoring matrix (PSSM) is introduced as its descriptor, which is generated by using the PSI-BLAST program[53] to search the NCBI's NR database(ftp://ftp.ncbi.nih.gov/blast/db/nr) through three iterations and a cutoff E-value 0.001 for multiple sequence alignment against the protein sequence P. The PSSM is a matrix of size $L \times 20$, where L is the length of the query amino acid sequence and 20 represent the 20 native amino acid types. The sample of a protein P can be represented by the following equation:

$$P_{PSSM} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{pmatrix} \tag{2.2}$$

where $P_{i,j}$ represents the score of the amino acid residue in the ith position of the protein sequence being changed to amino acid type j in the biology evolution process. In this work, the PSSM elements are mapped to the range of [0,1] using a standard sigmoid

function:

$$f(x) = 1/(1 + e^{-x}). \tag{2.3}$$

where x is the original PSSM value.

2.2.2 Three different autocorrelation descriptors based on PSSM

With the help of the knowledge of stochastic process, a protein sequence can be viewed as a time sequence of the corresponding physicochemical properties. In this study, only the evolutionary information represented in the form of PSSM is adopted as the considered properties. Here, each column is taken as one property , so the PSSM contains 20 different properties, which can be considered as the time sequences of all properties.

To transform the PSSM of different lengths into equal length vector, one approach[19] is to represent a protein sample P by

$$\overline{P}_{PSSM} = (\overline{P}_1, \overline{P}_2, \cdots, \overline{P}_{20})^T \tag{2.4}$$

where

$$\overline{P}_j = \frac{1}{L}\sum_{i=1}^{L} P_{i,j} \quad (j = 1, 2, \cdots, 20) \tag{2.5}$$

where $\overline{P}_j$ represents the average score of the amino acid residues in the protein P being mutated to amino acid type j during the evolution process. However, if $\overline{P}_{PSSM}$ (denoted by Ave-PSSM in this study) is only used to represent the protein P, all the sequence-order information during the evolution process would be lost. Hence, three different autocorrelation descriptors based on PSSM are adopted, which include normalized Moreau-Broto autocorrelation[54], Moran autocorrelation[55] and Geary autocorrelation[56].

Autocorrelation descriptor is a powerful statistical tool and defined based on the distribution of amino acid properties along the sequence, which measures the correlation between two residues separated by a distance of d in terms of their evolution scores, and they are defined as:

a) Normalized Moreau-Broto autocorrelation descriptors

$$N_j^d = \frac{1}{L-d}\sum_{i=1}^{L-d} P_{i,j} \times P_{i+d,j}, (j = 1, 2, \cdots, 20; d < L, d \neq 0) \tag{2.6}$$

where N_j^d is the Moreau-Broto correlation factor of amino acid type j, d is the lag of the autocorrelation along the protein sequence, $P_{i,j}$ and $P_{i+d,j}$ represents the score values in

the ith and $i+d$th position of the protein sequence being mutated to amino acid type j during the evolution process. The parameter d must be smaller than the length of the shortest sequence in the datasets. In this paper, the length of the shortest sequence for our datasets is 10 (1189 dataset), hence the value of d varies from 1 to 9. Then, a MBAC-PSSM feature vector is defined by combined normalized Moreau-Broto autocorrelation features with $\overline{P}_{PSSM}$ (denoted by MBAC-PSSM), and would be expressed as follows:

$$P^d_{MBACP} = (N^d_1, N^d_2, \cdots, N^d_{20})^T, \quad (d = 1, 2, \cdots, 9) \tag{2.7}$$

$$P_{MBACP} = (\overline{P}_{PSSM}, P^1_{MBACP}, P^2_{MBACP}, \cdots, P^9_{MBACP})^T \tag{2.8}$$

and the dimension of P_{MBACP} is 200.

b) Moran autocorrelation descriptors

$$M^d_j = \frac{\frac{1}{L-d}\sum_{i=1}^{L-d}(P_{i,j} - \overline{P_j})(P_{i+d,j} - \overline{P_j})}{\frac{1}{L}\sum_{i=1}^{L}(P_{i,j} - \overline{P_j})^2}, (j = 1, 2, \cdots, 20; d < L, d \neq 0) \tag{2.9}$$

where M^d_j is the Moran correlation factor of amino acid type j, d, $P_{i,j}$ and $P_{i+d,j}$ are the same as the above. $\overline{P}_j$ represents the average score of the amino acid residues in the protein P being mutated to amino acid type j during the evolution process. Similarity, a MAC-PSSM feature vector is defined by

$$P^d_{MACP} = (M^d_1, M^d_2, \cdots, M^d_{20})^T, \quad (d = 1, 2, \cdots, 9) \tag{2.10}$$

$$P_{MACP} = (\overline{P}_{PSSM}, P^1_{MACP}, P^2_{MACP}, \cdots, P^9_{MACP})^T \tag{2.11}$$

and the dimension of P_{MACP} is 200.

c) Geary autocorrelation descriptors

$$G^d_j = \frac{\frac{1}{2(L-d)}\sum_{i=1}^{L-d}(P_{i,j} - P_{i+d,j})^2}{\frac{1}{L-1}\sum_{i=1}^{L}(P_{i,j} - \overline{P_j})^2}, (j = 1, 2, \cdots, 20; d < L, d \neq 0) \tag{2.12}$$

where G^d_j is the Geary correlation factor by coupling the dth-most contiguous PSSM scores along the protein chain for the amino acid type j, $\overline{P}_j$, $P_{i,j}$ and $P_{i+d,j}$ are the same as the above. Then, a GAC-PSSM feature vector is defined by

$$P^d_{GACP} = (G^d_1, G^d_2, \cdots, G^d_{20})^T, \quad (d = 1, 2, \cdots, 9) \tag{2.13}$$

$$P_{GACP} = (\overline{P}_{PSSM}, P^1_{GACP}, P^2_{GACP}, \cdots, P^9_{GACP})^T \tag{2.14}$$

and the dimension of P_{GACP} is 200.

To cover more information, we propose a comprehensive model called MBMGAC-PSSM by fusing the 20 average score features selected from PSSM, the 180 normalized Moreau-Broto autocorrelation features, the 180 Moran autocorrelation features and the 180 Geary autocorrelation features. Finally, each protein sequence is characterized by a 560-dimensional feature vector:

$$P = (\psi_1, \psi_2, \cdots, \psi_\mu, \cdots, \psi_{560})^T \tag{2.15}$$

2.3 Feature selection

The dimension of our constructed feature vector is 560, which is a large input for SVM. The large dimension will lead three problems: over-fitting, information redundancy or noise and dimension disaster. Hence, feature selection plays a key role in classification task. Principal component analysis (PCA)[22, 31] is one effective dimensionality reduction methods. The goal of PCA is to select some dominant features which can retain most of the information in terms of an orthogonal transformation.

Let $X = (x_1, x_2, \cdots, x_t, \cdots, x_N)$ be a set of N input samples, each sample is a m-dimensional feature vector $x_t = (x_{t1}, x_{t2}, \cdots, x_{tm})^T$. PCA first solves an eigenvalue problem, assume that S is the sample covariance matrix of X:

$$S = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu)(x_t - \mu)^T \tag{2.16}$$

where μ is the sample mean, so that:

$$\lambda_i u_i = S u_i \qquad (i = 1, 2, \cdots, m) \tag{2.17}$$

where $\lambda_1, \lambda_2 \cdots, \lambda_m$ is the corresponding eigenvalue of the eigenvector $u_1, u_2 \cdots, u_m$ of S. Let λ_i be sorted in descending order, so λ_i is the ith largest eigenvalue. PCA adopts the corresponding eigenvectors of the first n eigenvalues to project the original samples into a n-dimensional orthogonal space using the linear transformation as follows:

$$Y = U^T X \tag{2.18}$$

where U^T is a $n \times m (n < m)$ matrix, which consists of the eigenvectors of the first n eigenvalues. Each feature vector of the samples Y in the new orthogonal space is viewed as a principal component.

In this work, our method is designed based on the 1189 dataset, then PCA is employed for the 1092 samples, each of which has 560 features, then the 175 features are obtained in the orthogonal space to perform the protein structural classes prediction.

2.4 Support vector machine

Support vector machine (SVM)[57] is a class of supervised machine learning algorithms based on statistical learning theory. SVM mainly is used to deal with statistical classification and regression analysis. Owing to the ability of condensing information contained in the training set, SVM often achieves outstanding classification performance, and it has been broadly applied in prediction of protein structural classes[31, 32, 58]. The basic idea of SVM is to find the separating hyperplane based on the support vector theory to minimize classification errors. It transforms the input data of samples to a higher dimensional space using the kernel function to find support vectors. Generally, four kinds of kernel functions, i.e. linear function, polynomial function, sigmoid function and radial basis function (RBF), can be available to perform prediction. In this study, we choose the RBF as SVM's kernel, which is defined as $\mathrm{K}(x, x^{'})$=$\exp(-\gamma\|x - x^{'}\|^2)$. The kernel parameter γ in addition to the regularization parameter C are optimized based on the training set (1189 dataset) by fifteen-fold cross validation using a grid search strategy in the LIBSVM package[59, 60].

A grid search strategy is a systematic testing of an entire range of values for a set of n parameters. These parameter values are determined by dividing the range of interest of each parameter into equal segments. Thus an initial range of values must be specified as well as the number of values to be examined for each parameter. The grid search strategy then procedes by examining all possible combinations of these parameter values and stores that combination which comes closest to meeting the design criterion. Here, we determine the values of C and γ by aiming to achieve the highest overall prediction accuracy as possible. For this purpose, a simple grid search strategy is adopted, where C is allowed to take a value only between 2^{-5} to 2^{15} and γ only between 2^{-15} to 2^{5}. By the above grid search, various pairs of (C; γ) values are tried and the one with the best cross-validation accuracy is selected.

2.5 Performance evaluation

In statistical prediction, cross-validation methods can be categorized into three types: independent dataset test, sub-sampling test and jackknife test. Among these three methods, the jackknife test is deemed the most rigorous and objective due to its ability of yielding a unique result for a given dataset. It has thus been increasingly and widely

used by investigators to examine the performance of various predictors[18, 19, 31–33]. Accordingly, we adopt jackknife test in this paper. During the process of the jackknife test, one protein sequence is singled out from the training set and the SVM classification model is trained by the remaining protein sequences. Then, the classification model is used to predict the singled out sequence. This process is repeated until every sequence in the training set has been singled out once. In this sense, the jackknife test is also known as the leave-one-out test.

To evaluate the performance of our method comprehensively, we report seven standard performance measures, including Sensitivity (Sens), Specificity (Spec), F-measure, Matthew's correlation coefficient (MCC), Area Under ROC Curve (AUC), Overall accuracy (OA) and Average accuracy(AA). F-value measures the performance of a test which is the harmonic mean of recall and precision. MCC represents the correlation coefficients between the observed and the predicted class. It's values ranges from +1 (indicating best prediction model) to -1 (indicating worst prediction model). The ROC analysis usually applies to binary classification problems. One of the classes is selected as a "positive" one. The ROC chart plots the true positive rate as a function of the false positive rate. It is parameterized by the probability threshold values. The true positive rate represents the fraction of positive cases that were correctly classified by the model. The false positive rate represents the fraction of negative cases that were incorrectly classified as positive. Each point on the ROC plot represents a (true positive rate/false positive rate) pair corresponding to a particular probability threshold. AUC is the area calculated under receiver operating characteristic (ROC) curve plotted by FP rate vs TP rate. It's values ranges from 0 to 1. These measures are defined by the following formulas:

$$Recall \ or \ Sens = \frac{TP}{TP+FN} \tag{2.19}$$

$$Spec = \frac{TN}{FP+TN} \tag{2.20}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.21}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.22}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.23}$$

$$AUC = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}) \tag{2.24}$$

$$OA = \frac{TP+TN}{TP+FN+FP+TN} \tag{2.25}$$

$$AA = \sum \frac{Sens}{n} \tag{2.26}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives and FN represents the number of false negatives, n represents the number of classes. To provide an intuitive picture, the general architecture of our proposed feature extraction method is shown in Figure 1.

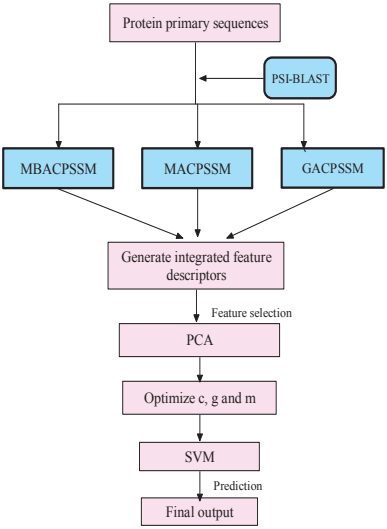

Figure 1. The general architecture of the proposed method.

3 Results and discussion

To predict protein structural class, a 560-dimensional feature vector is obtained. After the process of PCA, the dimension is reduced to 175 to avoid information redundancy, and then the 175 features are input into SVM. The RBF kernel function, the grid-search approach and fifteen-fold cross-validation for 1189 dataset are used to find the best parameters of C and γ for SVM. Finally, the optimal value of C and γ are computed to be 2 and 0.0019531.

3.1 Prediction performance of our model

The overall protein structural class prediction accuracy as well as the prediction accuracy for each structural class have achieved by using the combination of our features from the three submodels, which include MBAC-PSSM, MAC-PSSM and GAC-PSSM. The proposed prediction model(MBMGAC-PSSM) is examined with 1189, 25PDB and 640 datasets by jackknife tests and we report the Sens, Spec, F-measure, MCC and AUC for each structural class, the OA, as well as the AA. As shown in Table 1, relying solely on PSSM for feature extraction, we achieve up to 76.3%, 77.2% and 79.1% overall accuracies for 1189, 25PDB and 640 benchmarks, respectively, and average accuracies (AA) are also above 75.0% for three datasets. After comparing the four structural classes to each other, the values of Sens, Spec, F-measure, MCC and AUC in the all-α class, all-β class and α/β class are obviously separately superior to those of $\alpha+\beta$ class. However, referring to the $\alpha+\beta$ class, the prediction accuracy is relatively low compared with the other classes (only 50.6%, 61.7% and 63.2% for 1189, 25PDB and 640 datasets, respectively). The fact indicates that recognizing $\alpha+\beta$ class from the others is a challenging task due to its non-negligible overlap with the other classes. To improve its prediction accuracy, there are still many difficulties to overcome in the future study.

Table 1. The prediction quality of our model on the 1189, 25PDB and 640 datasets

Dataset	Structural class	Sens(%)	Spec(%)	F-measure	MCC	AUC
1189	All-α	79.8	94.0	0.79	0.73	0.87
	All-β	85.0	93.4	0.84	0.78	0.89
	α/β	84.7	89.4	0.81	0.73	0.87
	$\alpha+\beta$	50.6	91.3	0.56	0.45	0.71
	OA	76.3				
	AA	75.0				
25PDB	All-α	86.7	93.7	0.85	0.79	0.90
	All-β	81.5	93.5	0.82	0.75	0.87
	α/β	79.5	93.4	0.78	0.72	0.86
	$\alpha+\beta$	61.7	89.0	0.64	0.52	0.75
	OA	77.2				
	AA	77.4				
640	All-α	86.2	97.8	0.89	0.86	0.92
	All-β	83.1	94.0	0.82	0.77	0.89
	α/β	85.3	91.6	0.82	0.75	0.88
	$\alpha+\beta$	63.2	88.3	0.65	0.52	0.76
	OA	79.1				
	AA	79.5				

3.2 Prediction performance of our submodels

To explore the impact of our submodels on the protein structural class prediction accuracy, one by one, we add the other features groups to the Ave-PSSM features. From Table 2, we note that the prediction accuracy using 20 Ave-PSSM features only reaches 68.0%, 65.8% and 64.6% for 1189, 25PDB and 640 datasets, respectively. By adding MBAC, MAC and GAC features to 20 Ave-PSSM features, respectively, we achieve an improvement more than 6.0% for 1189 dataset, more than 8.0% for 25PDB dataset and more than 9.5% for 640 dataset. Three autocorrelation descriptors defined on PSSM do reflect intrinsic correlation and make their positive contributions and improvement to the overall predictions. Then, we combine MAC features with MBAC-PSSM features and build up a new submodel called MBAC-MAC-PSSM. By doing series of experiments, we achieve up to 76.2%, 76.7% and 77.0% prediction accuracy respectively for 1189, 25PDB and 640 datasets, which is 0.8%, 1.8% and 2.8% higher than that given only by MBAC-PSSM, and 0.1%, 0.5% and 2.1% lower than that obtained by MBMGAC-PSSM. As we can see, each submodel has played a positive role in improving the protein structural class prediction accuracy.

Table 2. Performance comparison of our submodels on three datasets.

Dataset	Features	Prediction accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	Ave-PSSM	72.7	78.2	80.8	33.6	68.0
	MBAC-PSSM	78.0	86.4	85.0	46.1	75.4
	MAC-PSSM	81.6	86.7	81.1	41.5	74.0
	GAC-PSSM	80.3	85.7	82.6	41.9	74.0
	MBAC-MAC-PSSM	81.6	87.1	81.4	50.6	76.2
	MBMGAC-PSSM	79.8	85.0	84.7	50.6	76.3
25PDB	Ave-PSSM	78.6	69.8	67.3	47.9	65.8
	MBAC-PSSM	84.7	81.0	75.1	58.7	74.9
	MAC-PSSM	88.7	80.1	76.0	57.1	75.5
	GAC-PSSM	85.8	80.1	75.7	55.1	74.1
	MBAC-MAC-PSSM	87.6	82.2	75.1	61.7	76.7
	MBMGAC-PSSM	86.7	81.5	79.5	61.7	77.2
640	Ave-PSSM	65.2	63.6	79.7	48.5	64.4
	MBAC-PSSM	78.3	79.2	87.0	53.2	74.2
	MAC-PSSM	84.8	83.8	83.0	58.5	77.0
	GAC-PSSM	82.6	84.4	84.2	57.9	76.9
	MBAC-MAC-PSSM	80.4	83.1	84.8	60.1	77.0
	MBMGAC-PSSM	86.2	83.1	85.3	63.2	79.1

Furthermore, the receiver operating characteristic (ROC) curves on three submodels are implemented to evaluate the prediction performance for the different submodels. Figure 2 shows the ROC curves for the 25PDB dataset by this method and the other three

models (including MBAC-PSSM, MAC-PSSM and GAC-PSSM). The area under curve (AUC) of this method is 0.915, which is higher than those by MBAC-PSSM, MAC-PSSM and GAC-PSSM individually (AUCs are 0.895, 0.904 and 0.897, respectively). Similar results are obtained for the other two datasets (figures are not shown). This further indicates that MBMGAC-PSSM is more effective for improving the prediction of protein structural. Meanwhile, we can see that the submodel MAC-PSSM can give more information than the other submodels.

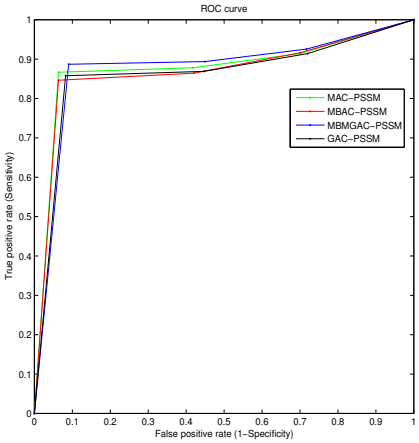

Figure 2. ROC curves of different submodels on the 25PDB dataset.

3.3 Performance comparison with other methods

In this section, to demonstrate the superiority of our model, the proposed method is further compared with the other recently reported prediction methods on the same datasets. We select the accuracy of each class and overall accuracy as evaluation indexes that are shown in Table 3. The compared methods include the famous methods SCPRED [61] and MODAS[58], SCPRED mainly based on the information extracted from the predicted protein secondary structure sequence, MODAS combines evolutionary profiles and predicted secondary structure. Generally speaking, the prediction accuracies of these methods that contain predicted protein secondary structure information are higher than that only based on evolutionary information. Hence, SCPRED and MODAS are listed in Table 3 only as two reference methods. AAD-CGR[62] is proposed to analyze amino acids sequence by

recurrence quantification analysis based on chaos game representation, whose prediction accuracy is only 65.2% and 64.0% for 1189 and 25PDB datasets, respectively. SCEC[50] incorporates evolutionary information encoded using PSI-BLAST profile-based collocation of AA pairs, whose prediction accuracy is 9.6% and 16.8% lower than that of our model for 25PDB and 640 datasets, respectively. The compared methods also include other competitive methods such as RPSSM[32], AADP-PSSM[18], AAC-PSSM-AC[19], AATP[31] and MEDP[33] are recently reported protein structural classes prediction methods based on the evolutionary information represented in the form of PSSM. RPSSM is a submodel from PSSS-PSSM[32]. As can be seen from Table 3, among five PSSM-based

Table 3. Performance comparison of different methods on three datasets.

Dataset	Method	Prediction accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	SCPRED[61]	89.1	86.7	89.6	53.8	80.6
	MODAS[58]	92.3	87.1	87.9	65.4	83.5
	RPSSM[32]	67.7	75.2	74.6	17.4	60.2
	AAD-CGR[62]	62.3	67.7	66.5	63.1	65.2
	AADP-PSSM[18]	69.1	83.7	85.6	35.7	70.7
	AATP[31]	72.7	85.4	82.9	42.7	72.6
	MEDP[33]	85.2	84.0	84.3	45.2	75.8
	AAC-PSSM-AC[19]	80.7	86.4	81.4	45.2	74.6
	This paper	**79.8**	**85.0**	**84.7**	**50.6**	**76.3**
25PDB	SCPRED[61]	92.6	80.1	74.0	71.0	79.7
	MODAS[58]	92.3	83.7	81.2	68.3	81.4
	SCEC[50]	75.8	75.2	82.6	31.8	67.6
	RPSSM[32]	75.6	70.2	52.0	43.3	60.8
	AAD-CGR[62]	64.3	65.0	65.0	61.7	64.0
	AADP-PSSM[18]	83.3	78.1	76.3	54.4	72.9
	AATP[31]	81.9	74.7	75.1	55.8	71.7
	MEDP[33]	87.8	78.3	76.0	57.4	74.8
	AAC-PSSM-AC[19]	85.3	81.7	73.7	55.3	74.1
	This paper	**86.7**	**81.5**	**79.5**	**61.7**	**77.2**
640	SCPRED[61]	90.6	81.8	85.9	66.7	80.8
	SCEC[50]	73.9	61.0	81.9	33.9	62.3
	MEDP[33]	84.8	75.3	86.4	53.8	74.7
	This paper	**86.2**	**83.1**	**85.3**	**63.2**	**79.1**

methods, our model achieves the highest overall prediction accuracy with improvement of 0.5-16.1%, 2.4-16.4% and 4.4% for 1189, 25PDB and 640 datasets, respectively. The overall accuracies are 0.5%, 2.4% and 4.4% higher than the previous best-performing results that are obtained by the MEDP model. For 1189 dataset, although all-α, all-β and α/β classes accuracies are not the highest, our model still obtain the satisfactory results. Refer to $\alpha+\beta$ class, our model achieves relatively high result, the accuracy reaches 50.6%

with improvement of 5.4-33.2%. As for the 25PDB dataset, the prediction accuracy of all-α class and all-β class is only 1.1% and 0.2% lower than the highest value from MEDP and AAC-PSSM-AC, respectively. However, for α/β and $\alpha+\beta$ classes, we obtain the best results, which is 3.2% and 4.3% higher than that given by AATP-PSSM and MEDP, respectively. For 640 dataset, except α/β class, the prediction accuracies of other three classes are higher than those obtained by MEDP. Obviously, our proposed model has a great improvement for the prediction accuracy of $\alpha+\beta$ class, which indicates that our proposed model reflects some critical information related to the $\alpha+\beta$ class due to the usage of three autocorrelation descriptors.

4 Conclusions

In this study, the main contribution is to construct a 560-dimensional feature vector by defining three autocorrelation descriptors: normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation on PSSM, which measure the correlation between two residues separated by a distance of d in terms of their evolution scores. Then 175 features are selected using PCA. The SVM classifier and the jackknife test are employed to predict and evaluate the model on three benchmark datasets: 1189, 25PDB and 640 datasets, with sequence similarity lower than 40% , 25% and 25%, respectively. The experiment results show that our proposed method is very promising and may provide a cost-effective alternative to predict protein structural class in particular for low-similarity datasets. We shall make efforts in our future work to provide a public accessible web-server for the method presented in this paper. The codes used to prepare this paper are available from the author upon request.

Acknowledgements: The authors thank the anonymous referees for their many valuable suggestions that have improved this manuscript. This work was supported by the National Natural Science Foundation of China (Nos. 61373174, 11326201) and the Fundamental Research Funds for the Central Universities (No. JB140703).

References

[1] K. C. Chou, Progress in protein structural class prediction and its impact to bioinformatics and proteomics, *Curr. Protein Pept. Sci.* **6** (2005) 423–436.

[2] M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, *Protein Eng.* **11** (1998) 249–251.

[3] C. T. Zhang, Z. Zhang, Z. He, Prediction of the secondary structure contents of globular proteins based on three structural classes, *J. Protein Chem.* **17** (1998) 261–272.

[4] Z. Zhang, Z. R. Sun, C. T. Zhang, A new approach to predict the helix/strand content of globular proteins, *J. Theor. Biol.* **208** (2001) 65–78.

[5] L. Carlacci, K. C. Chou, G. M. Maggiora, A heuristic approach to predicting the tertiary structure of bovine somatotropin, *Biochemistry* **30** (1991) 4389–4398.

[6] A. Anand, G. Pugalenthi, P. N. Suganthan, Predicting protein structural class by SVM with class-wise optimized features and decision probabilities, *J. Theor. Biol.* **253** (2008) 375–380.

[7] M. Levitt, C. Chothia, Structural patterns in globular proteins, *Nature* **261** (1976) 552–557.

[8] G. P. Zhou, An intriguing controversy over protein structural class prediction, *J. Protein Chem.* **17** (1998) 729–738.

[9] K. C. Chou, A key driving force in determination of protein structural classes, *Biochem. Biophys. Res. Commun.* **264** (1999) 216–224.

[10] Y. D. Cai, G. P. Zhou, Prediction of protein structural classes by neural network, *Biochimie* **82** (2000) 783–785.

[11] Y. D. Cai , X. J. Liu, X. B. Xu, K. C. Chou, Prediction of protein structural classes by support vector machines, *J. Comput. Chem.* **26** (2002) 293–296.

[12] T. L. Zhang, Y. S. Ding, Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes, *Amino Acids* **33** (2007) 623–629.

[13] X. Xiao, S. H. Shao, Z. D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, *J. Comput. Chem.* **27** (2006) 478–482.

[14] T. L. Zhang, Y. S. Ding, K. C. Chou, Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern, *J. Theor. Biol.* **250** (2008) 186–193.

[15] R. Y. Luo, Z. P. Feng, J. K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, *Eur. J. Biochem.* **269** (2002) 4219–4225.

[16] X. D. Sun, R. B. Huang, Prediction of protein structural classes using support vector machines, *Amino Acids* **30** (2006) 469–475.

[17] K. C. Chou, Y. D. Cai, Predicting protein structural class by functional domain composition, *Biochem. Biophys. Res. Commun.* **321** (2004) 1007–1009.

[18] T. G. Liu, X. Q. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, *Biochimie* **92** (2010) 1330–1334.

[19] T. G. Liu, X. B. Geng, X. Q. Zheng, R. Li, J. Wang, Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, *Amino Acids* **42** (2012) 2243–2249.

[20] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, *J. Theor. Biol.* **267** (2010) 272–275.

[21] S. L. Zhang, S. Y. Ding, T. M. Wang, High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure, *Biochimie* **93** (2011) 710–714.

[22] Z. C. Li, X. B. Zhou, Z. Dai, X. Y. Zou, Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, *Amino Acids* **37** (2009) 415–425.

[23] L. Li, X. Cui, S. Yu, Y. Zhang, Z. Luo, H. Yang, Y. Zhou, X. Zheng, PSSP-RFE: Accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical–chemical property and function annotations, *PloS One* **9** (2014) e92863.

[24] S. Y. Ding, S. J. Yan, S. H. Qi, Y. Li, Y. H. Yao, A protein structural classes prediction method based on PSI-BLAST profile. *J. Theor. Biol.* **353** (2014) 19–23.

[25] D. Cai, G. P. Zhou, Prediction of protein structural classes by neural network, *Biochimie* **82** (2000) 783–785.

[26] C. Chen, Y. X. Tian, X. Y. Zou, P. X. Cai, J. Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, *J. Theor. Biol.* **243** (2006) 444–448.

[27] D. Cai, X. J. Liu, X. B. Xu, K. C. Chou, Prediction of protein structural classes by support vector machines, *J. Comput. Chem.* **26** (2002) 293–296.

[28] H. B. Shen, J. Yang, X. J. Liu, K. C. Chou, Using supervised fuzzy clustering to predict protein structural classes, *Biochem. Biophys. Res. Commun.* **334** (2005) 577–581.

[29] Z. X. Wang, Z. Yuan, How good is prediction of protein structural class by the component-coupled method? *Proteins* **38** (2000) 165–175.

[30] Y. F. Cao, S. Liu, L. D. Zhang, J. Qin, J. Wang, K. X. Tang, Prediction of protein structural class with rough sets, *BMC Bioinform.* **7** (2006) 20.

[31] S. L. Zhang, Y. Feng, X. G. Yuan, Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, *J. Biomol. Struct. Dyn.* **29** (2012) 634–642.

[32] S. Y. Ding, Y. Li, Z. X. Shi, S. J. Yan, A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, *Biochimie* **97** (2014) 60–65.

[33] L. C. Zhang, X. Q. Zhao, L. Kong, Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition, *J. Theor. Biol.* **355** (2014) 105–110.

[34] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen and Y. Z. Chen, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Res.* **34** (2006) W32–W37.

[35] C. Chen, L. X. Chen, X. Y. Zou, P. X. Cai, Predicting protein structural class based on multi-features fusion, *J. Theor. Biol.* **253** (2008) 388–392.

[36] W. Chen, P. M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, *Nucleic Acids Res.* **41** (2013) e68.

[37] Y. Xu, J. Ding, L. Y. Wu, iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, *PloS One* **8** (2013) e55844.

[38] W. Chen, P. M. Feng, E. Z. Deng, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **462** (2014) 76–83.

[39] H. Lin, E. Z. Deng, H. Ding, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **42** (2014) 12961–12972.

[40] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PloS One* **9** (2014) e106691.

[41] Y. Xu, X. Wen, L. S. Wen, L. Y. Wu, iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PloS One* **9** (2014) e105018.

[42] W. R. Qiu, X. Xiao, W. Z. Lin, iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, *J. Biomol. Struct. Dyn.* (2014) in press.

[43] X. Xiao, J. L. Min, W. Z. Lin, Z. Liu, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *J. Biomol. Struct. Dyn.* (2014) in press.

[44] Q. Dai, Y. Li, X. Liu, Y. Yao, Y. Cao, P. He, Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position, *BMC Bioinform.* **14** (2013) #152 (pp. 1–14).

[45] Q. Dai, W. Li, L. Li, Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features, *J. Comput. Chem.* **32** (2011) 3393–3398.

[46] J. Wang, Y. Li, X. Liu, Q. Dai, Y. Yao, P. He, High-accuracy prediction of protein structural classes using PseAA structural properties and secondary structural patterns, *Biochimie* **101** (2014) 104–112.

[47] J. Wang, C. Wang, J. Cao, X. Liu, Y. Yao, Q. Dai, Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features, *Gene* **554** (2015) 241–248.

[48] S. L. Zhang, Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC, *Chemometr. Intell. Lab.* in press.

[49] L. A. Kurgan, L. Homaeian, 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** (2006) 2323–2343.

[50] K. Chen, L. A. Kurgan, J. S. Ruan, Prediction of protein structural class using novel evolutionary collocation-based sequence representation, *J. Comput. Chem.* **29** (2008) 1596–1604.

[51] K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), *J. Theor. Biol.* **273** (2011) 236–247.

[52] K. C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, *Proteins: Struct. Funct. Genet.* **43** (2001) 246–255.

[53] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25** (1997) 3389–3402.

[54] G. Moreau, P. Broto, Autocorrelation of molecular structures, application to SAR studies, *Nour. J. Chim.* **4** (1980) 757–764.

[55] P. A. Moran, Notes on continuous stochastic phenomena, *Biometrika* **37** (1950) 17–23.

[56] R. C. Geary, The contiguity ratio and statistical mapping. *Incorp. Statistician* **5** (1954) 115–145.

[57] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[58] M. J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, *BMC Bioinform.* **10** (2009) #414 (pp. 1–24).

[59] C. C. Chang, C. J. Lin. LIBSVM: a library for support vector machines, (2001).

[60] http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[61] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, *BMC Bioinform.* **9** (2008) #226 (pp. 1–15).

[62] J. Y. Yang, Z. L. Peng, Z. G. Yu, R. J. Zhang, V. Anh, D. S. Wang, Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *J. Theor. Biol.* **257** (2009) 618–626.