Chemometrics and Intelligent Laboratory Systems 142 (2015) 28–35

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC

Shengli Zhang [*]

School of Mathematics and Statistics, Xidian University, Xi'an 710071, PR China

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 5 January 2015
Accepted 9 January 2015
Available online 17 January 2015

Keywords:
Protein structural class prediction
Secondary structure sequences
Support vector machine
Transition probability matrix
Pseudo-position specific scoring matrix

ABSTRACT

Determination of protein structural class using a fast and suitable computational method has become a hot issue in protein science. Prediction of protein structural class for low-similarity sequences remains a challenge problem. In this study, a 111-dimensional feature vector is constructed to predict protein structural classes. Among the 111 features, 100 features based on pseudo-position specific scoring matrix (PsePSSM) are selected to reflect the evolutionary information and the sequence-order information, and the other 11 rational features based on predicted protein secondary structure sequences (PSSS) are designed in the previous works. To evaluate the performance of the proposed method (named by PSSS–PsePSSM), jackknife cross-validation tests are performed on three widely used benchmark datasets: 1189, *25PDB* and 640. Our method achieves competitive performance on prediction accuracies, especially for the overall prediction accuracies for datasets 1189, *25PDB* and 640, which reach 86.6%, 89.5% and 81.0%, respectively. The PSSS–PsePSSM algorithm also outperforms other existing methods, indicating that our proposed method is a cost-effective computational tool for protein structural class prediction.

1. Introduction

Knowledge of protein structure plays a key role in analysis of protein functions, protein binding, rational drug design, and many other related fields and applications. The structure is organized on several levels, which include the primary, secondary, tertiary, and quaternary structures. The knowledge of structural classes of proteins is very important and useful in many aspects of molecular biology [1], especially for improving the accuracy of secondary structure prediction [2] and reducing the search space of the possible conformations for the tertiary structure [3,4]. According to the definition proposed by Levitt and Chothia [5] in 1976, proteins can be classified into the following four structural classes: all-α, all-β, α/β and $\alpha + \beta$ classes. The structural class is one of the important attributes for a protein, which plays an important role in both experimental and theoretical studies in protein science.

During the past two decades, many efforts had been made to predict the protein structural class. Since Nakashima et al. [6] indicated that the protein structural class is related to its amino acid composition (AAC), most of the existing prediction methods are proposed by using the simple sequence representation such as composition vectors [7–11]. However, representing a protein sample solely with its AAC, many important features associated with the sequence orders are completely missed. To avoid losing much important information hidden in protein sequences, various descriptors containing the sequence-order effects are proposed for enhancing the prediction quality, including pair-coupled amino acid composition [12], polypeptide composition [13,14], pseudo-amino acid composition (PseAAC) [15], and functional domain composition [16].

Although the promising results have been achieved in many cases, the existing methods appear to be less effective in low-homology datasets, whose average pairwise sequence identities are less than 40%. In order to improve the prediction accuracy for low-homology proteins, several novel features on the basis of the predicted protein secondary structure sequences and position-specific scoring matrix (PSSM) are developed. In SCPRED [17], the input of the classifier includes 8 features based on information extracted from the predicted secondary structure with PSI-PRED and the other feature computed from the amino acid sequence. In RKS-PPSC [18], 24 features are designed by using recurrence quantification analysis, *k*-string based information entropy and segment-based analysis. In MODAS [19] and PSSS–PSSM [20], the predicted secondary structure information is employed to perform the prediction with evolutionary information. The AADP-PSSM [21] method extends the traditional dipeptide composition to PSSM. The AATP model [22] fuses AAC and transition probability composition from PSSM, and the AAC–PSSM-AC [23] model combines auto covariance and PSSM to extract evolutionary information. Although the existing methods have shown satisfactory performance for low-similarity datasets, there is always space for improvement.

[*] Tel./fax: +86 29 88202860.
E-mail address: shengli0201@163.com.

http://dx.doi.org/10.1016/j.chemolab.2015.01.004
0169-7439/© 2015 Elsevier B.V. All rights reserved.

Compared to the conventional AAC, the PseAAC firstly introduced by Chou [15] can include the sequence-order information. Since the concept of pseudo amino acid composition [15] or Chou's PseAAC [25–27] is proposed, it has penetrated into almost all the areas of computational proteomics (see a long list of references cited in [28]) as well as many biomedicine and drug development areas [29]. Since it has been widely and increasingly used, recently three powerful open access softwares, called 'PseAAC-Builder' [25], 'propy' [26], and 'PseAAC-General' [28], have been established: the former two are for generating various modes of Chou's special PseAAC including the predicted secondary structure sequence or PSSS [54]; while the third one for those of Chou's general PseAAC, including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs. (9)–(10) of [38]), "Gene Ontology" mode (see Eqs. (11)–(12) of [38]), and "Sequential Evolution" or "PSSM" mode (see Eqs. (13)–(14) of [38]). Similarly, the Pseudo-PSSM (PsePSSM) is also proposed by Shen and Chou [24] in order to incorporate the evolution information and the sequence-order information. In this study, we propose a new comprehensive method (called PSSS–PsePSSM) by fusing the 100 features from PsePSSM and the 11 existing features from the predicted secondary structure sequences. Jackknife cross-validation tests on three widely used benchmark datasets show that our method achieves the satisfactory performance in comparison with the other existing methods, particularly for the low-similarity amino acid sequences.

As demonstrated by a series of recent publications [30–37] in response to the call [38], to establish a really useful sequence-based statistical predictor for a biological system, we need to follow the following 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) introduce or develop a powerful algorithm (or engine) 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

To be comparable with the previous works, three popular benchmark datasets are used to evaluate the performance of our method: the 1189 dataset [39], the *25PDB* dataset [40] and the 640 dataset [71], 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 contains 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 vectors

To develop a powerful predictor for a protein system, one of the keys is to effectively define feature vectors to formulate the statistical samples concerned. According to Eq.(6) of [38], the feature vector for any protein, peptide or biological sequence is just the general form of pseudo-amino acid composition or PseAAC [15] that can be formulated as

$$P = \left(\psi_1, \psi_2, \cdots, \psi_\mu, \cdots, \psi_\Omega\right)^T \qquad (1)$$

where T is the transpose operator, while the subscript Ω is an integer and its value as well as the components $\psi_1, \psi_2, \cdots, \psi_\Omega$ will depend on how to extract the desired information from the amino acid sequence of P.

In this section, we will explain the PSSS–PsePSSM feature extraction strategies.

2.2.1. Features based on the predicted protein secondary structure sequence

According to the structural classification of proteins (SCOP), the protein domains are generally categorized into the four major structural classes: (1) all-α class, which is essentially formed by helices and only includes small amount of strands, (2) all-β class, which is essentially formed by strands and only includes small amount of helices, (3) α/β class, which includes both the helices and mostly parallel strands, and (4) $\alpha+\beta$ class, which includes both the helices and mostly antiparallel strands. It is well known that every amino acid in a protein sequence can be predicted into one of the three secondary structural elements: H(helix), E(strand), and C(coil). This study uses PSIPRED [41], which predicts protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST [42]. In this work, the 11-dimensional feature vectors based on the predicted secondary structural sequence have been constructed in the previous works [17,43–45], which can be expressed as:

$$P_{PSSS} = (p_1, p_2, \cdots, p_{11})^T \qquad (2)$$

a. p_1 and p_2 represent the occurrences of the secondary structural elements $H(Con_H)$ and $E(Con_E)$, respectively, in the secondary structural sequence. The sequence length is denoted by N.
b. p_3 and p_4 represent normalized length of the longest α-helices ($MaxSeg_H/N$) and β-strands ($MaxSeg_E/N$), respectively.
c. p_5 and p_6 represent normalized average length of α-helices ($AvgSeg_H/N$) and β-strands ($AvgSeg_E/N$), respectively.
d. Two composition moment features, $p_7(CMV_H)$ and $p_8(CMV_E)$, which are formulated as

$$CMV_H = \frac{\sum_{j=1}^{n_H} n_{Hj}}{N(N-1)}, \qquad CMV_E = \frac{\sum_{j=1}^{n_E} n_{Ej}}{N(N-1)} \qquad (3)$$

where n_H and n_E are the total number of H and E in the sequence of the secondary structure, respectively; n_{Hj} and n_{Ej} are the jth position (in the secondary structure sequence) of H and E, respectively.
e. The ultimate three features [45] from the secondary structure sequences characterizing the distributions of α helices and β strands, and they are designed to improve the prediction accuracies of the $\alpha+\beta$ and α/β classes. Firstly, a secondary structure sequence is converted into a segment sequence, which is composed of *helix segments* and *strand segments* (denoted by α and β, respectively). Here, a helix (strand) segment refers to a continuous segment of all $H(E)$ symbols in the secondary structure sequence. Then, in order to concentrate on the arrangement of α helix and β strand segments, the coil segments are ignored in the reduced segment sequence.

The transition probability matrix (*TPM*) [45] of the reduced segment sequence can be defined as follows:

$$TPM = \begin{pmatrix} P_{\alpha\alpha} & P_{\alpha\beta} \\ P_{\beta\alpha} & P_{\beta\beta} \end{pmatrix} \qquad (4)$$

In order to measure the degree of segment aggregation, $p_9(P_{\alpha\beta})$ and $p_{10}(P_{\beta\alpha})$ are selected to add into our feature set, where $P_{\alpha\beta} = N_{\alpha\beta}/(N_{\alpha\beta} + N_{\alpha\alpha})$, $P_{\beta\alpha} = N_{\beta\alpha}/(N_{\beta\alpha} + N_{\beta\beta})$, $N_{\alpha\alpha}$, $N_{\alpha\beta}$, $N_{\beta\alpha}$ and $N_{\beta\beta}$ enumerate the content of substring $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$ and $\beta\beta$ in a segment sequence, respectively. The third feature to be extracted is the probability of strand segments occurring in a segment sequence, denoted by $p_{11}(P(\beta))$.

2.2.2. Features based on the pseudo-position specific scoring matrix

To reflect the evolutionary information, we utilize each protein sequence as a seed to search and align homogenous sequences from NCBI's NR database (ftp://ftp.ncbi.nih.gov/blast/db/nr)using the PSI-BLAST program [42] with three iterations and a cutoff E-value 0.001, then the PSSM is constructed through a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is a log-odds matrix of size $L \times 20$, where L is the length of the query amino acid sequence and 20 is due to the 20 amino acids. The sample of a protein P can be represented by

$$P_{PSSM} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{pmatrix} \tag{5}$$

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 during the evolution process. In this work, we transform the PSSM elements to [0,1] using the following sigmoid function:

$$f(x) = 1/\left(1 + e^{-x}\right), \tag{6}$$

where x is the original PSSM value.

To make the PSSM descriptor become a size-uniform matrix, one approach is to represent a protein sample P by

$$\overline{P}_{PSSM} = \left(\overline{P}_1, \overline{P}_2, \cdots, \overline{P}_{20}\right)^T, \tag{7}$$

where

$$\overline{P}_j = \frac{1}{L}\sum_{i=1}^{L} P_{i,j} \quad (j = 1, 2, \cdots, 20) \tag{8}$$

where $\overline{P}_j$ is the composition of the amino acid type j in the PSSM and represents the average score of the amino acid residues in the protein P being mutated to amino acid type j during the evolution process. $\overline{P}_{PSSM}$ is denoted by PSSM–AAC [23]. However, if $\overline{P}_{PSSM}$ is only used to represent the protein P, all the sequence-order information during the evolution process will be lost. To reflect the sequence-order information, we adopt the concept of PsePSSM [24] and obtain the PsePSSM features according to the following equations [46]:

$$P^{\lambda}_{PsePSSM} = \left(\theta^{\lambda}_1, \theta^{\lambda}_2, \cdots, \theta^{\lambda}_j, \cdots, \theta^{\lambda}_{20}\right)^T \tag{9}$$

$$\theta^{\lambda}_j = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\left(P_{i,j} - P_{(i+\lambda),j}\right)^2, (j = 1, 2, \cdots, 20; \lambda < L, \lambda \neq 0) \tag{10}$$

where θ^{λ}_j is the correlation factor of amino acid type j, whose contiguous distance is λ along the protein sequence. Then, the PsePSSM feature vector can be expressed as follows:

$$P_{PsePSSM} = \left(\overline{P}_1, \overline{P}_2, \cdots, \overline{P}_{20}, \theta^1_1, \theta^1_2, \cdots, \theta^1_{20}, \theta^2_1, \theta^2_2, \cdots, \theta^2_{20}, \cdots, \theta^{\lambda}_1, \theta^{\lambda}_2, \cdots, \theta^{\lambda}_{20}\right)^T. \tag{11}$$

and the dimension of $P_{PsePSSM}$ is $20 + (20 \times \lambda)$.

To cover more effective information, we propose a novel computational model called PSSS–PsePSSM by fusing the $20 + (20 \times \lambda)$ features selected from PsePSSM and the predicted secondary structure sequence.

2.3. Support vector machine

Support vector machine (SVM) is a machine learning algorithm based on Vapnik's [47] statistical learning theory, which can be used for classification, regression analysis, sorting and so on. Among these applications, classification is the fundamental task for SVM and has been broadly applied in prediction of protein structural classes. The basic idea of SVM is to map data of samples into a high dimensional Hilbert space and to explore an optimal separating hyperplane in this space. The mapping is determined by kernel functions. 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. Empirical studies have demonstrated that the RBF outperforms the other three kinds of kernel functions [48,49]. Hence, we choose the RBF to perform prediction, which is defined as $\mathrm{K}(x, x^{'}) = \exp(-\gamma \|x - x^{'}\|^2)$. The regularization parameter C and kernel parameter γ are optimized based on the training set (1189 dataset) by fifteenfold cross validation using a grid search strategy in the LIBSVM software [50,51].

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 proceeds 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 $(C; \gamma)$ values are tried and the one with the best cross-validation accuracy is selected.

2.4. Prediction assessment

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling test, and jackknife test [60]. However, as elucidated in [38] and demonstrated by Eqs. (28)–(32) of [38], among the three cross-validation methods, the jackknife test is deemed the least arbitrary (most objective) that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (see, e.g., [46, 52–55,61–70]). Accordingly, the jackknife test is also adopted here to examine the quality of the present predictor. During the process of test, each protein sequence in the samples is singled out in turn as a test sample and the remaining protein sequences are used as training samples.

Additionally, we have reported the performance of our approach using the following qualified measures: the individual sensitivity (*Sens*), individual specificity (*Spec*), overall prediction accuracy (*OA*), average prediction accuracy (*AA*), *F*-measure and Matthew's correlation coefficient (*MCC*). These measures can be defined as:

$$Precision = \frac{TP}{TP + FP}(P) \tag{12}$$

$$Recall\ or\ sensitivity = \frac{TP}{TP + FN}(R,\ Sens) \tag{13}$$

$$Specificity = \frac{TN}{TN + FP}(Spec) \tag{14}$$

$$Overall\ accuracy = \frac{TP + TN}{TP + FN + FP + TN}(OA) \tag{15}$$

$$Average\ \ accuracy\ =\ \frac{\sum Sens}{\xi}(AA) \tag{16}$$

$$F{-}measure\ =\ \frac{2\times P\times R}{P+R} \tag{17}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{18}$$

where *TP* is the number of true positives, *FP* is the number of false positives, *TN* is the number of true negatives and *FN* is the number of false negatives, respectively. *F*-value measures the performance of a test which is the harmonic mean of recall and precision. The *MCC* value takes account of both over- and under-predictions and is between −1 and 1. ξ denotes the number of classes.

According to the formulation proposed recently in [33,34,56], the above metrics can be also expressed in another form (see Eq. (11) of [56] or Eq. (16) of [34], or Eq. (11) of [33]). As we can see from the above literatures, it is much more intuitive and easier-to-understand when using these expressions to examine a predictor for its sensitivity, specificity, overall prediction accuracy and *MCC*, particularly for its *MCC*. Also, it is instructive to point out that the set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology [57] and system medicine [58], a completely different set of metrics as defined in [59] is needed. To provide an intuitive picture, the general framework of our proposed feature extraction method is shown in Fig. 1.

3. Results and discussion

3.1. Selecting the optimal parameter λ

The parameter λ of PsePSSM represents the distance between the two considered amino acid residues in the sequence, the maximal λ must be smaller than the length of the shortest sequence in the dataset. In this paper, the length of the shortest sequence for *25PDB* dataset is 13, and for 1189 dataset it is 10, hence the value of the parameter λ varies from 1 to 9. Through a series of control experiments on our datasets, as shown in Fig. 2, the value of λ is optimized as 4 for the PSSS–PsePSSM model on 1189, *25PDB* and 640 datasets. We also detect that the value of λ has minor influence on the overall prediction accuracy with the increase of λ, indicating that our model is reliable and robust. Here, we are to use a combination of PSSS and PsePSSM to represent the protein samples, and $\Omega = 111$. Finally, a 111-dimensional feature set is constructed and applied to perform the protein structural class prediction. The optimal values of *C* and λ are computed to be 5.6569 and 1 based on 1189 dataset, which are used in the following experiments to avoid over-fitting problem.

3.2. Prediction performances of our method

The proposed prediction method is examined with 1189, *25PDB* and 640 datasets in low similarity by jackknife test and report the Sensitivity, Specificity, *MCC* and *F*-measure for each structural class, as well as the *OA* and *AA*. From Tables 1–2, the results show that the overall accuracies for the two datasets are all above 86%, which reach 86.6% and 89.5% for the datasets 1189 and *25PDB*, respectively. Table 3 shows

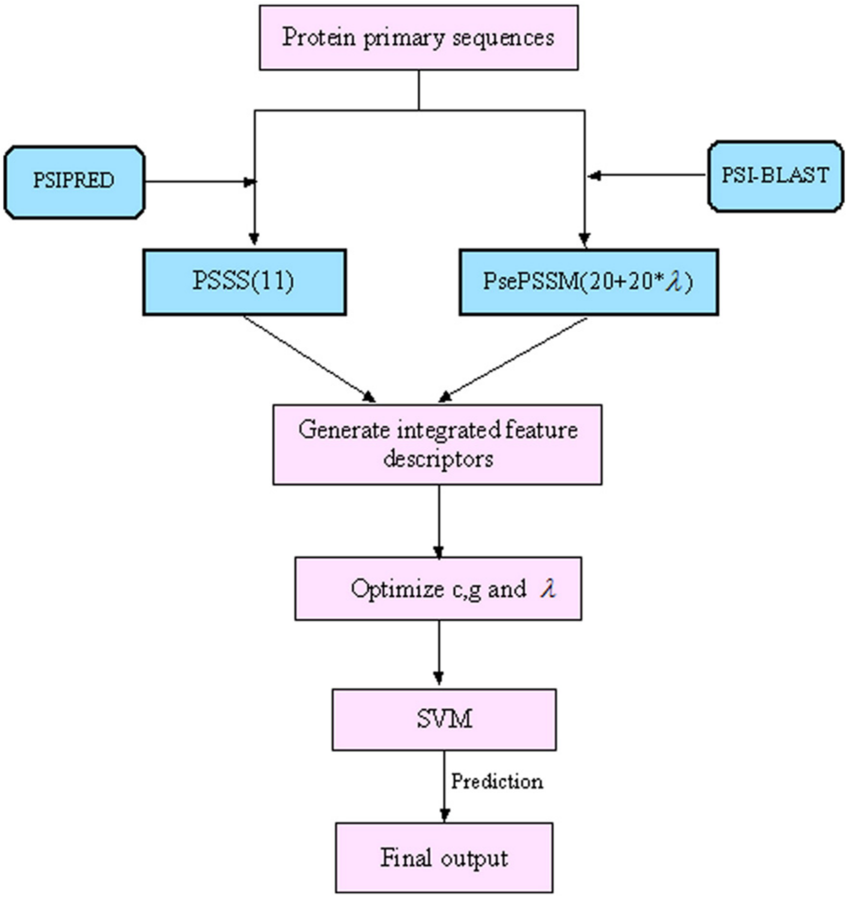

Fig. 1. The general framework of the proposed PSSS–PsePSSM model.

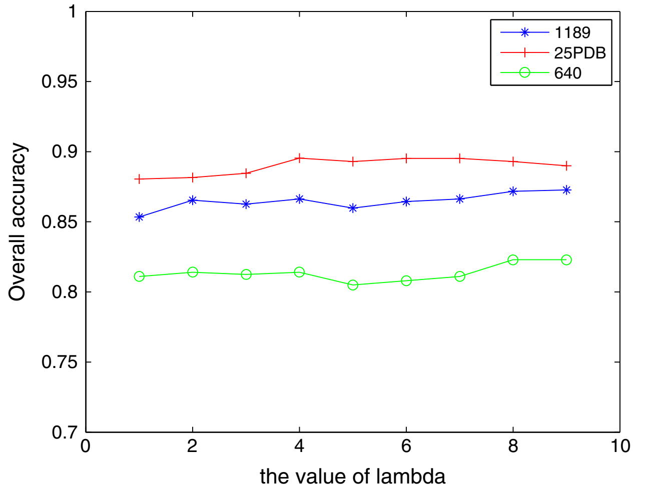

Fig. 2. The overall accuracies of different values of λ for the PSSS–PsePSSM model on the *25PDB*, 1189 and 640 datasets.

that the overall accuracy for the 640 dataset is 81.0%. The average accuracies (*AA*) are also above 80.0% for three datasets. After comparing the four structural classes to each other, the predictions of proteins with sensitivities, specificities, *F*-measure and *MCC* in the all-α class and all-β class are better than the other classes, and they are above 80% for all the datasets. Referring to the α/β class, our method also performs satisfactorily with prediction accuracies of about 87.7%, 90.2% and 84.7%, respectively. However, it seems very difficult to predict the $\alpha + \beta$ class, and their prediction accuracies are relatively low compared with the other classes, especially for 640 dataset (only 70.8%). The low prediction accuracy of the $\alpha + \beta$ class may be due to its non-negligible overlap with the other classes [71].

In this study, the 11 PSSS features are rationally utilized to reflect the general content and spatial arrangement of the secondary structural elements of a given protein sequence, while the other 20 PsePSSM features (PSSM–AAC) are extracted to reflect the evolutionary information, and the 80 PsePSSM features mainly reflect the sequence-order information. Table 4 shows the performance comparison of different choices of features. Referring to Table 4, the overall accuracy obtained by PSSS is about 6.9–10.8% higher than that obtained by PsePSSM for three datasets. For 1189 and *25PDB* datasets, with the addition of PSSM–AAC features to PSSS, the overall accuracies are 2.9% and 3.3% higher than those of the PSSS model, indicating that the evolutionary information do help to improve the accuracies. Similarly, with the addition of PsePSSM features to PSSS, the overall accuracies have an improvement of 5.3% and 6.3%, respectively. As for 640 dataset, the PSSS–PsePSSM model performs satisfactorily although the overall accuracy is relatively low compared with that of the PSSS model, this may be due to the influences of information redundancy and noise. We also notice that the PSSS–PsePSSM model achieves the best overall accuracy which is 2.4%, 3.0% and 0.6% higher than that of the PSSS–PSSM–AAC model, respectively. The fact reveals that the incorporation of sequence-order information does make their positive contributions and improvement to the overall predictions.

Table 1
The prediction quality of our method on the 1189 dataset.

Structural class	Sens(%)	Spec(%)	*MCC*(%)	F-measure(%)	AUC(%)
All-α	91.9	97.7	89.3	91.5	94.8
All-β	91.8	98.1	90.9	93.3	95.0
α/β	87.7	93.4	80.5	86.6	90.6
$\alpha + \beta$	73.9	92.8	66.9	74.2	83.4
OA	86.6				
AA	86.3				

Table 2
The prediction quality of our method on the *25PDB* dataset.

Structural class	Sens(%)	Spec(%)	*MCC*(%)	F-measure(%)	AUC(%)
All-α	96.4	98.3	94.3	95.9	97.3
All-β	90.5	97.0	87.8	91.0	93.8
α/β	90.2	96.9	86.4	89.3	93.5
$\alpha + \beta$	81.2	93.8	75.4	81.8	87.5
OA	89.5				
AA	89.6				

3.3. Comparison of accuracies between different classification algorithms

This work selects the SVM as the classifier. To have a comparative performance study, the proposed method is analyzed with many well studied classifiers such as K-nearest neighbor algorithm (KNN) with $K = 9$ and linear discriminant analysis (LDA). The success rates of all the classifiers are evaluated with all the three benchmark datasets (1189, *25PDB* and 640 datasets). The overall classification accuracies as well as the accuracies for each structural class are presented in Table 5. Table 5 shows that the SVM performs the best among the three classifiers. The average predicted accuracy of the SVM for the three datasets is 85.7%, which is 1.9–7.8% higher than that of other two classifiers.

Furthermore, the receiver operating characteristic (ROC) curves on three datasets are implemented to evaluate the prediction performance for the different classification algorithms. Fig. 3 shows the ROC curves for the *25PDB* dataset by this method and the other two algorithms. The area under curve (AUC) of this method is 0.93, which is higher than those by KNN and LDA individually (AUCs are 0.84 and 0.88, respectively). Similar results are obtained for the other two datasets (figures are not shown). This further indicates that the SVM is more suitable for prediction of protein structural classes based on PSSS and PsePSSM.

3.4. Comparison with existing methods

In this section, the proposed method is further compared with other recently reported prediction methods on the same datasets. We select the accuracy of each class and overall accuracy as evaluation indexes which are summarized in Table 6. The compared methods include the famous methods SCPRED [17] and MODAS [19], and the other competitive methods such as RKS-PPSC [18], IEA-PSSF [72], Zhang et al. [45], PSSS–PSSM [20], LCC-PSSM [73], AADP-PSSM [21] and AAC-PSSM-AC [23]. Among these methods, the PSSS–PSSM method has the best overall accuracy on the three datasets. The SCPRED, RKS-PPSC, IEA-PSSF and Zhang et al. methods are mainly based on the information extracted from the predicted protein secondary structure sequence. The MODAS and PSSS–PSSM models combine the predicted protein secondary structure information and evolutionary profile to perform prediction. The AADP-PSSM, LCC-PSSM and AAC–PSSM-AC models are recently reported protein structural class prediction methods based on the evolutionary information extracted from PSSM.

As shown in Table 6, our method achieves the highest overall prediction accuracies for 1189, *25PDB* and 640 among all the compared methods, which reach 86.6%, 89.5% and 81.0%, respectively. The overall

Table 3
The prediction quality of our method on the 640 dataset.

Structural class	Sens(%)	Spec(%)	*MCC*(%)	F-measure(%)	AUC(%)
All-α	87.0	97.8	86.4	89.2	92.4
All-β	81.2	96.7	80.3	84.8	88.9
α/β	84.7	92.4	76.2	82.9	88.6
$\alpha + \beta$	70.8	86.8	56.3	68.4	78.8
OA	81.0				
AA	80.9				

Table 4
Performance comparison of different choices of features.

Dataset	Features	Prediction accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	PSSS	91.0	86.1	78.4	70.5	81.3
	PSSM-AAC	65.0	77.9	84.1	24.1	65.3
	PSSS + PSSM-AAC	90.6	87.4	84.7	73.4	84.2
	PsePSSM	82.0	82.3	84.1	44.0	74.4
	PSSS + PsePSSM	91.9	91.8	87.7	73.9	86.6
25PDB	PSSS	92.8	81.7	82.4	75.7	83.2
	PSSM-AAC	79.9	67.7	66.5	45.8	64.9
	PSSS + PSSM-AAC	93.5	86.5	86.7	79.6	86.5
	PsePSSM	86.2	78.8	75.7	57.6	75.5
	PSSS + PsePSSM	96.4	90.5	90.2	81.2	89.5
640	PSSS	89.1	79.2	88.1	74.3	82.5
	PSSM-AAC	56.5	58.4	76.8	29.2	55.3
	PSSS + PSSM-AAC	85.5	79.8	83.0	74.2	80.4
	PsePSSM	73.9	76.6	85.3	51.5	71.7
	PSSS + PsePSSM	87.0	81.2	84.7	70.8	81.0

accuracies are 1.4%, 2.9% and 0.2% higher than the previous best-performing results that are obtained with the IEA-PSSF method, the PSSS–PSSM method and the SCPRED method, respectively. Especially for the *25PDB* dataset, the proposed method also achieves the highest all-β class, α/β class, $\alpha+\beta$ class prediction accuracies. For the all-α class, the prediction accuracy is only 0.2% lower than the highest value from PSSS-PSSM method. Further, comparing the accuracy values with those obtained by the famous method MODAS and SCPRED, there are improvements of 8.1% and 9.8%, respectively. As for the 1189 dataset, our method achieves the highest all-β class prediction accuracies. The all-α class and α/β class, our method also obtains favorable prediction accuracy although it is not the highest one. Compared with MODAS and SCPRED about the overall accuracies, there are also improvements of 3.1% and 6.0%. We also notice that the significant improvement is made in particular for the β class, which is the challenging class to predict. For example, the proposed method obtains the 81.2% and 70.8% accuracy of the $\alpha+\beta$ class for the datasets *25PDB* and 640, which is 2.3% and 4.1% higher than that given with PSSS–PSSM and SCPRED, respectively. The dataset 1189 obtains 73.9%, which is 0.4% lower than that given by IEA-PSSF method, but still obtains the satisfactory results among the other prediction methods.

Meanwhile, PsePSSM includes not only the evolutionary information but also the sequence-order information. The fact can be taken into consideration to further improve the prediction accuracy. Although sequence similarity on the 1189 dataset is higher than that on the *25PDB* dataset, the overall prediction accuracy for the *25PDB* dataset is higher than that for the 1189 dataset. This further illustrates that our proposed method is effective and powerful for the low-similarity datasets. In summary, the outstanding performance of the current method can be attributed to the effective usage of the features that are extracted from the secondary structure predicted with PSIPRED as well as that from PsePSSM.

Table 5
Comparison of accuracies between different classification algorithms.

Dataset	Algorithms	Prediction accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	KNN	85.7	87.4	87.4	45.6	77.8
	LDA	91.9	90.5	87.7	74.3	86.4
	SVM	**91.9**	**91.8**	**87.7**	**73.9**	**86.6**
25PDB	KNN	89.2	84.9	92.2	53.3	79.2
	LDA	92.6	87.8	85.0	73.9	84.8
	SVM	**96.4**	**90.5**	**90.2**	**81.2**	**89.5**
640	KNN	83.3	83.1	93.8	47.4	76.6
	LDA	84.1	85.7	85.3	67.3	80.3
	SVM	**87.0**	**81.2**	**84.7**	**70.8**	**81.0**

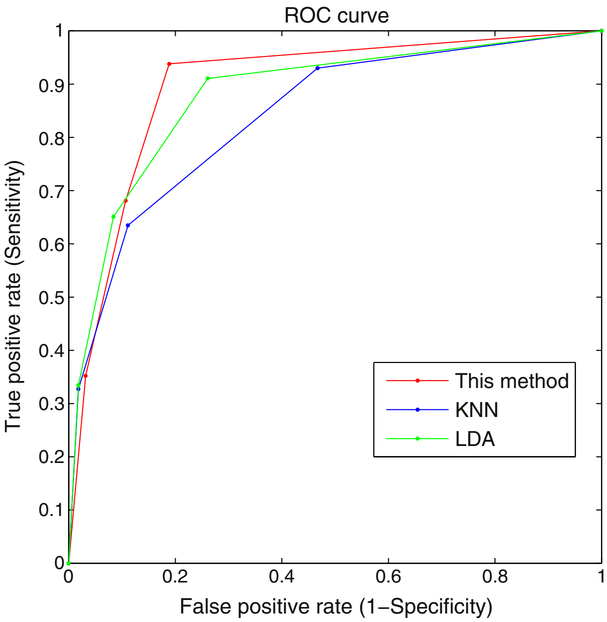

Fig. 3. ROC curves of different methods on the *25PDB* dataset.

4. Conclusions

Accurate prediction of protein structural classes for low-similarity sequences is a complicated and challenging task in the current computational biology. Though some of the existing methods have shown the state-of-the-art performance, space for improvement remains. In this study, the main contribution is to propose a new comprehensive feature set, which includes the 11 features that are rationally utilized to reflect the general contents and spatial arrangements of the secondary structural elements of a given protein sequence, and the other 100 features are extracted based on the evolutionary information and the sequence-order information from PsePSSM. Three widely used datasets 1189, *25PDB* and 640, with sequence similarity lower than 40%, 25% and

Table 6
Performance comparison of different methods on three datasets.

Dataset	Method	Prediction accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	SCPRED [17]	89.1	86.7	**89.6**	53.8	80.6
	MODAS [19]	92.3	87.1	87.9	65.4	83.5
	RKS-PPSC [18]	89.2	86.7	82.6	65.6	81.3
	IEA-PSSF [72]	94.2	91.5	81.4	**74.3**	85.2
	AADP-PSSM [21]	69.1	83.7	85.6	35.7	70.7
	Zhang et al. [45]	92.4	87.4	82.0	71.0	83.2
	PSSS-PSSM [20]	**94.2**	88.4	85.3	71.8	85.0
	LCC-PSSM [73]	89.2	88.8	85.6	58.5	81.2
	AAC-PSSM-AC [23]	80.7	86.4	81.4	45.2	74.6
	This paper	91.9	**91.8**	87.7	73.9	**86.6**
25PDB	SCPRED [17]	92.6	80.1	74.0	71.0	79.7
	MODAS [19]	92.3	83.7	81.2	68.3	81.4
	RKS-PPSC [18]	92.8	83.3	85.8	70.1	82.9
	IEA-PSSF [72]	90.1	84.7	79.5	77.6	83.1
	AADP-PSSM [21]	83.3	78.1	76.3	54.4	72.9
	Zhang et al. [45]	95.0	85.6	81.5	73.2	83.9
	PSSS-PSSM [20]	**96.6**	87.1	83.0	78.9	86.6
	LCC-PSSM [73]	91.7	80.8	79.8	64.0	79.0
	AAC-PSSM-AC [23]	85.3	81.7	73.7	55.3	74.1
	This paper	96.4	**90.5**	**90.2**	**81.2**	**89.5**
640	SCPRED [17]	**90.6**	**81.8**	85.9	66.7	80.8
	SCEC [71]	73.9	61.0	81.9	33.9	62.3
	MEDP [75]	84.8	75.3	**86.4**	53.8	74.7
	This paper	87.0	81.2	84.7	**70.8**	**81.0**

The accuracies are evaluated by jackknife test and measured by the percentage of correctly predicted proteins. The best results are highlighted in bold face.

25%, respectively, are adopted to assess the performance of our method. Results by jackknife tests show that our proposed method is competitive and can be used as the potential candidate for the accurate prediction of protein structural classes.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [74], we shall make efforts in our future work to provide a public accessible web-server for the method presented in this paper. When the web-server is established, we'll make an announcement via the CHEMOLAB Journal. The codes used to prepare this paper are available from the author upon request.

Conflict of interest

No conflict of interest.

Acknowledgments

The author thanks the anonymous referees for their many valuable suggestions that have improved this manuscript. This work was supported by the TianYuan Special Fund of the National Natural Science Foundation of China (grant no. 11326201), the Fundamental Research Funds for the Central Universities (no. JB140703), and the Natural Science Basic Research Plan in Shaanxi Province of China.

References

[1] L.A. Kurgan, T. Zhang, H. Zhang, S. Shen, J. Ruan, Secondary structure-based assignment of the protein structural classes, Amino Acids 35 (2008) 551–564.
[2] M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, Protein Eng. 11 (1998) 249–251.
[3] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[4] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Understanding the recognition of protein structural classes by amino acid composition, Proteins 29 (1997) 172–185.
[5] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–557.
[6] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 152–162.
[7] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.
[8] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216–224.
[9] Y.D. Cai, G.P. Zhou, Prediction of protein structural classes by neural network, Biochimie 82 (2000) 783–785.
[10] Y.D. Cai, X.J. Liu, X.B. Yu, G.P. Zhou, Support vector machines for predicting protein structural class, BMC Bioinforma. 2 (2001) 1–5.
[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] K.C. Chou, Using pair-coupled amino acid composition to predict protein secondary structure content, J. Protein Chem. 18 (1999) 473–480.
[13] 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.
[14] X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.
[15] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255.
[16] K.C. Chou, Y.D. Cai, Predicting protein structural class by functional domain composition, Biochem. Biophys. Res. Commun. 321 (2004) 1007–1009.
[17] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, BMC Bioinforma. 9 (2008) 226.
[18] J. Yang, Z. Peng, X. Chen, Prediction of protein structural classes for low-homology sequences based on predicted secondary structure, BMC Bioinforma. 11 (2010) S9.
[19] M.J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, BMC Bioinforma. 10 (2009) 414.
[20] 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.
[21] T. Liu, X. 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.
[22] 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.
[23] T.G. Liu, X.B. Geng, X.Q. Zheng, et al., Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, Amino Acids 42 (2012) 2243–2249.
[24] H.B. Shen, K.C. Chou, NUC-PLOC: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM, Protein Eng. Des. Sel. 20 (2007) 561–567.
[25] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2012) 117–119.
[26] D.S. Cao, Q.S. Xu, Y.Z. Liang, propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.
[27] S.X. Lin, J. Lapointe, Theoretical and experimental biology in one, J. Biomed. Sci. Eng. 6 (2013) 435–442.
[28] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.
[29] W.Z. Zhong, S.F. Zhou, Molecular science for drug development and biomedicine, Int. J. Mol. Sci. 15 (2014) 20072–20078.
[30] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[31] 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.
[32] 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.
[33] 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.
[34] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, iDNA-Protdis: 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.
[35] 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.
[36] 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) 968875, http://dx.doi.org/10.1080/07391102.
[37] 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), http://dx.doi.org/10.1080/07391102.07392014.07998710.
[38] 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.
[39] Z.X. Wang, Z. Yuan, How good is prediction of protein structural class by the component-coupled method? Proteins 38 (2000) 165–175.
[40] K.D. Kedarisetti, L. Kurgan, S. Dick, Classifier ensembles for protein structural class prediction with varying homology, Biochem. Biophys. Res. Commun. 348 (2006) 981–988.
[41] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[42] 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.
[43] 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 Recognit. 39 (2006) 2323–2343.
[44] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, J. Theor. Biol. 267 (2010) 272–275.
[45] 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.
[46] G.L. Fan, Q.Z. Li, Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 334 (2013) 45–51.
[47] V. Vapnik, The Nature of Statistical Learning Theory, 1st ed. Springer, NY, 1995.
[48] Z. Yuan, B. Huang, Prediction of protein accessible surface areas by support vector regression, Proteins 57 (2004) 558–564.
[49] Z. Yuan, T.L. Bailey, R.D. Teasdak, Prediction of protein B-factor profiles, Proteins 58 (2005) 905–912.
[50] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2001.
[51] http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[52] K.C. Chou, H.B. Shen, Recent progress in protein subcelluar localization prediction, Anal. Biochem. 370 (2007) 1–16.
[53] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, Biochem. Biophys. Res. Commun. 337 (2005) 752–756.
[54] L. Kong, L.C. Zhang, J.F. Lv, Accurate prediction of protein structural classes by incorporating predicted secondary structural information into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 344 (2014) 12–18.
[55] M.K. Gupta, R. Niyogi, M. Misra, An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition, SAR QSAR Environ. Res. 24 (2013) 597–609.
[56] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.

[57] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2012) 629–641.

[58] X. Xiao, P. Wang, W.Z. Lin, J.H. Jia, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.

[59] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.

[60] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[61] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans. Comput. Biol. Bioinforma. 9 (2012) 467–475.

[62] H. Mohabatkar, M.M. Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach, Med. Chem. 9 (2013) 133–137.

[63] B. Liu, X. Wang, Q. Zou, Q. Dong, Q. Chen, Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation, Mol. Inform. 32 (2013) 775–782.

[64] D.N. Georgiou, T.E. Karakasidis, A.C. Megaritis, A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory, Open Bioinformatics J. 7 (2013) 41–48.

[65] X. Niu, L. Nana, X. Jingbo, C. Dingyan, P. Yuehua, X. Yang, W. Weiquan, W. Dongming, W. Zengzhen, Using the concept of Chou's pseudo amino acid composition to predict protein solubility: an approach with entropies in information theory, J. Theor. Biol. 332 (2013) 211–217.

[66] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.

[67] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2014) 197–203.

[68] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. 341 (2014) 34–40.

[69] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 365 (2014) 96–103.

[70] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition, J. Theor. Biol. 360 (2014) 109–116.

[71] 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.

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

[73] 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.

[74] K.C. Chou, H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 2 (2009) 63–92.

[75] 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.