Knowledge-Based Systems 163 (2019) 787–793

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Predicting protein structural classes for low-similarity sequences by evaluating different features

Xiao-Juan Zhu [a], Chao-Qin Feng [a], Hong-Yan Lai [a], Wei Chen [a,b,*], Lin Hao [a,*]

[a] *Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[b] *Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China*

HIGHLIGHTS

- A novel method is developed to predict protein structural classes.
- The protein samples were formulated by integrating various knowledge.
- An overall accuracy of 96.7% was obtained on a strict benchmark dataset.

ARTICLE INFO

Article history:
Received 2 March 2018
Received in revised form 2 September 2018
Accepted 4 October 2018
Available online 15 October 2018

Keywords:
Protein structural class
Feature fusion
Low-similarity sequence
Machine learning method

ABSTRACT

Protein structural class could provide important clues for understanding protein fold, evolution and function. However, it is still a challenging problem to accurately predict protein structural classes for low-similarity sequences. This paper was devoted to develop a powerful method to predict protein structural classes for low-similarity sequences. On the basis of a very objective and strict benchmark dataset, we firstly extracted optimal tripeptide compositions (OTC) which was picked out by using feature selection technique to formulate protein samples. And an overall accuracy of 91.1% was achieved in jackknife cross-validation. Subsequently, we investigated the accuracies of three popular features: position-specific scoring matrix (PSSM), predicted secondary structure information (PSSI) and the average chemical shift (ACS) for comparison. Finally, to further improve the prediction performance, we examined all combinations of the four kinds of features and achieved the maximum accuracy of 96.7% in jackknife cross-validation by combining OTC with ACS, demonstrating that the model is efficient and powerful. Our study will provide an important guide to extract valuable information from protein sequences.

1. Introduction

Protein structure plays an important role in understanding its function. According to their chain fold topologies, protein domains are generally categorized into four main structural classes: all-α, all-β, α/β and $\alpha+\beta$ proteins, which are shown in Fig. 1 [1]. All-α and all-β proteins are mainly composed of α-helices and β-strands, respectively. α/β proteins mainly consisted of α-helices and β-strands alternately with β-sheets almost built up from parallel strands. $\alpha+\beta$ proteins are predominantly made up of α-helices and β-strands separately with β-sheets almost formed by anti-parallel strands. α/β and $\alpha+\beta$ proteins are always combined as mixed $\alpha\beta$ proteins because the two classes are different in the aspect of the secondary structure connectivity, which is considered at a lower level describing topology [2]. The knowledge of protein structural class can effectively increase the accuracy of secondary structure and tertiary structure prediction [3,4], reduce the scope of conformational searches during energy optimization [5] and provide the important information about protein function [6–8].

In the past three decades, the prediction of protein structural classes has become one of the hotspots in bioinformatics and has attracted the attention of many bioinformatics scholars and structural biologists [9–35]. In the past ten years, many machine learning based algorithms have been used to build computational models for the prediction of protein structure classes, such as support vector machine (SVM) [9–15], artificial neural network (ANN) [16], covariant discriminant [17], Fisher's discriminant (FD) [18], increment of diversity combined with quadratic discriminant (IDQD) [19], Bayesian classifier [20,21], and so on. A key point for machine learning methods is to extract fixed-length and valid features to formulate protein samples. Hence, various sequence features have been applied to represent protein

[*] Corresponding authors.
E-mail addresses: zhuxiaojuan_uestc@163.com (X.-J. Zhu), f13551163107@163.com (C.-Q. Feng), laihongyan5386@163.com (H.-Y. Lai), chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (L. Hao).

https://doi.org/10.1016/j.knosys.2018.10.007
0950-7051/© 2018 Published by Elsevier B.V.

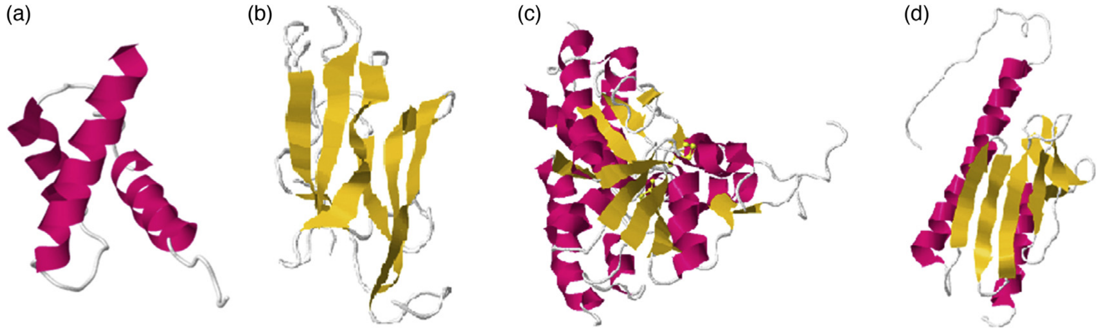

Fig. 1. A schematic illustration to show the four structural classes of proteins: (a) all-α (PDB ID: 1HQB), (b) all-β (PDB ID: 1S6N), (c) α/β (PDB ID: 1QXN) and (d) $\alpha+\beta$ (PDB ID: 2GA5).

sequences, such as amino acid composition (AAC) [22–25], dipeptide and tripeptide compositions [19,26,27], pseudo amino acid composition (PseAAC) [28], PSI-BLAST profile [29,30], predicted secondary structure information (PSSI) [31–33], average chemical shift (ACS) [11,34,35], and so on.

The prediction accuracy of these methods is strongly affected by the sequence similarity of the training and testing datasets. For example, the accuracy is more than 90% for high-similarity sequences [35–37]. However, it is less than 85% for low-similarity sequences. Therefore, many efforts have been made to improve the prediction accuracy for low-similarity datasets by selecting different sequence features and classification algorithms in recent years [30,33,38]. However, the accuracy of the recent methods is still far from satisfactory.

In our previous studies [11,39], ACS and optimal k-mer peptide composition were successfully applied in protein structural class prediction with low-similarity sequences. The aim of this study is to develop a high accuracy model for the prediction of protein structural class with low-similarity sequences.

The paper is organized as follows. A high quality benchmark dataset was firstly built. And then the feature encoding schemes that has been used in protein structural classes have been introduced. Subsequently, we examined the prediction performance of different features including ACS, PSSM, OTC, PSSI, and their fusion. Finally, a SVM based model was proposed and was utilized to perform discrimination on a low-similarity (~15%) protein benchmark dataset. High accuracy was obtained by using the jackknife cross-validation.

2. Material and methods

2.1. Database

The proteins file with the chemical shift values of nuclei $^{13}C_O$, ^{13}C, ^{13}C, $^{1}H_N$, ^{1}H and ^{15}N were obtained from re-referenced protein chemical shift database (RefDB) [40]. Their structural class types and sequences were obtained from Protein Data Bank (PDB) [41]. In order to get a reliable and high quality benchmark dataset, we excluded the proteins (1) whose structural class type were not annotated in PDB, (2) whose sequence identity is less than 50 residues, (3) whose CSs of <35% residues are not provided. Finally, PISCES program was utilized to remove the high similarity sequences by using sequence identity cutoff of 25% [42]. Through the rigorous filtration of the above steps, we finally obtained a low-similarity dataset which includes 124 all-α, 112 all-β, 163 mixed $\alpha\beta$ proteins. Among the 399 proteins, 395 proteins share less than 15% sequence identity, suggesting that the benchmark dataset is very strict. Such reliable and rigorous dataset provide us a strict and objective standard to evaluate the performances of various prediction methods.

2.2. Optimal tripeptide composition

To reflect the residue composition and their short correlation, the tripeptide composition [43] was produced by sliding a window of three residues with the step of one residue along a protein sequence **P**, which can be described as:

$$\mathbf{P} = [f_1, f_2, \ldots, f_i, \ldots, f_{8000}]^T \tag{1}$$

where the symbol $\boldsymbol{T}$ denotes the transposition of the vector, f_i represents the frequency of the ith tripeptide and can be expressed as:

$$f_i = n_i / \sum_{i=1}^{8000} n_i = n_i/(L-2) \tag{2}$$

where n_i and L denote the number of the ith tripeptide and the length of protein sequence, respectively.

However, an arbitrary tripeptide occurring in one type of protein structural classes is maybe a stochastic event. The tripeptide is redundant information or noise, which will bring out overfitting, overestimation and generation capability of the proposed model. Therefore, we must pick out the optimal tripeptides. Based on the statistical theory, the occurrence of a tripeptide in one type of protein structural classes obeys the binomial distribution [14]. Accordingly, the confidence level (CL) of the tripeptide i occurring in jth type can be calculated by

$$CL_{ij} = 1 - \sum_{k=n_{ij}}^{N_i} \frac{N_i!}{k!\,(N_i-k)!} q_j^k \left(1-q_j\right)^{N_i-k} \tag{3}$$

where n_{ij} is the number of the ith tripeptide in the jth type of protein. N_i is the number of the ith tripeptide in all protein samples. q_j is called the prior probability and can be calculated by

$$q_j = m_j/M \tag{4}$$

where m_j is the number of tripeptide in the jth type of protein and M is the total occurrence frequency of all tripeptides in the all protein samples. Since there are three types of protein structural classes in this study, each tripeptide has three CL values in the three types. We selected the maximum CL as the tripeptide's CL.

To find out the optimal tripeptides, the increment feature selection (IFS) [44,45] was used. At first, we ranked the 8000 tripeptides in a descending order according to their CL values. Subsequently, the first tripeptide with the largest CL value in the ranked feature set was regarded as the first feature subset. Then, the second feature subset was formed by adding the tripeptide with the second largest CL value into the first feature subset. This two-step process was repeated until all 8000 tripeptides were added. Finally, SVM was utilized to investigate the performances of the 8000 feature subsets by use of 5-fold cross-validation test for finding the optimal feature subset which can produce the maximum accuracy. As a result, we found that 1254 optimal tripeptides can produce the best prediction performance. The protein sequence **P** was thus formulated by optimal tripeptide composition (OTC) as follows.

$$\mathbf{P}_{\mathbf{OTC}} = [f_1, f_2, \ldots, f_i, \ldots, f_{1254}]^T \tag{5}$$

2.3. Position-specific score matrix (PSSM)

To reflect the evolutionary information, we utilized each protein sequence as a seed to search homogenous sequences from NCBI's NR database using the PSI-BLAST program [46] with three iterations and a cutoff E-value of 0.001. Then the PSSM was constructed through a multiple alignment of the highest scoring hits in an initial BLAST search.

PSSM is a matrix of size $L \times 20$, where L is the length of the protein primary sequence and there are 20 amino acids. The (i, j)th entry of the matrix represents the score of the residue in the ith position of query sequence being mutated to residue type j during the evolution process. In this work, the PSSM elements were scaled to the range from 0 to 1 using the following sigmoid function

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{6}$$

where x is the original PSSM value.

To make the PSSM descriptor become a size-uniform matrix, we extracted amino acid composition (AAC) and dipeptide composition (DPC) from the PSSM. A protein sample was represented by P_{PSSM}.

$$P_{PSSM} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix} \tag{7}$$

Then, an arbitrary protein **P** with L residues can be represented by AAC-PSSM as follows:

$$\mathbf{P} = (p_1, p_2, \ldots, p_j, \ldots, p_{20})^T \quad (j = 1, 2, \ldots, 20) \tag{8}$$

where

$$p_j = \frac{1}{L} \sum_{i=1}^{L} p_{i,j} \quad (j = 1, 2, \ldots., 20) \tag{9}$$

where p_j is the composition of residue type j in 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. However, AAC-PSSM signals only represent the residue composition in protein **P**, and all the sequence-order information will be lost. For reflecting the sequence-order information, the DPC-PSSM [30] was proposed to describe protein samples as follows:

$$\begin{aligned} \mathbf{P} = \big(&D_{1,1}, D_{1,2}, \ldots, D_{1,20}, D_{2,1}, D_{2,2}, \ldots, D_{2,20}, \ldots, \\ &D_{20,1}, D_{20,2}, \ldots, D_{20,20}\big)^T \end{aligned} \tag{10}$$

where

$$D_{i,j} = \frac{1}{L-1} \sum_{k=1}^{L-1} p_{k,i} \times p_{k+1,j} \quad (1 \le i, j \le 20) \tag{11}$$

where $p_{i,j}$ can be calculated by Eq. (7).

For including more information, AAC-PSSM and DPC-PSSM features are merged together into a 420-dimensional vector and can be expressed as:

$$\begin{aligned} \mathbf{P}_{\boldsymbol{PSSM}} = \big(&p_1, p_2, \ldots, p_L, D_{1,1}, D_{1,2}, \ldots, D_{1,20}, D_{2,1}, D_{2,2}, \ldots, \\ &D_{2,20}, \ldots, D_{20,1}, D_{20,2}, \ldots, D_{20,20}\big)^T \end{aligned} \tag{12}$$

2.4. Predicted protein secondary structure information

The protein secondary structure information is also important feature for protein structural class prediction. In this study, we used three secondary structure states namely: α-helix, β-strand and coil to describe the structure of residues in protein sequence. The software PSIPRED [47] were used to predict protein secondary structure. Then, each sample will be formulated by a 8-dimensional feature vector expressed as:

$$\mathbf{P}_{\boldsymbol{PSSI}} = (p_1, p_2, \ldots, p_8)^T \tag{13}$$

where p_1 and p_2 represent the content of the secondary structure α-helix and β-strand, respectively, and were formulated as:

$$\begin{cases} p_1 = n_\alpha / L \\ p_2 = n_\beta / L \end{cases} \tag{14}$$

where n_α and n_β are the numbers of residues belonging to α-helix and β-strand in the protein with the length of L residues.

The p_3 and p_4 denote the content of the longest α-helix segment and β-strand segment, respectively, and were formulated as::

$$\begin{cases} p_3 = n_{\alpha-max} / L \\ p_4 = n_{\beta-max} / L \end{cases} \tag{15}$$

where $n_{\alpha-max}$ and $n_{\beta-max}$ are the residue numbers of the longest α-helix segment and β-strand segment in the protein.

The p_5 and p_6, respectively represent the ratios of average lengths of α-helix and β-strand in the protein, and were formulated as:

$$\begin{cases} p_5 = \overline{n_\alpha} / L \\ p_6 = \overline{n_\beta} / L \end{cases} \tag{16}$$

where $\overline{n_\alpha}$ and $\overline{n_\beta}$ are the average lengths of α-helix segment and β-strand segment in the protein.

The p_7 and p_8 respectively represent composition moment vectors for α-helix and β-strand, which can reflect the information about both composition and position of residues in the sequence and are formulated as:

$$\begin{cases} p_7 = \dfrac{\sum_{j=1}^{n^\alpha} n_j^\alpha}{L(L-1)} \\ p_8 = \dfrac{\sum_{j=1}^{n^\beta} n_j^\beta}{L(L-1)} \end{cases} \tag{17}$$

where n^α and n^β are the total number of residues belonging to α-helix and β-strand, respectively. n_j^α and n_j^β are the jth position of α-helix and β-strand residue in the secondary structure sequence, respectively.

Most of the types of features used in this study can now be conveniently calculated and extracted using the state-of-the-art iFeature toolkit specifically developed for feature extraction of protein and peptide sequences [43].

2.5. Averaged chemical shift

The chemical shift (CS) is a measurement of nuclear magnetic resonance (NMR) to measure the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. Some works have demonstrated that the CS information is a powerful indicator for protein structure prediction [11,48,49]. Here, we calculated averaged chemical shifts (ACSs) as follows.

$$ACS_i = \sum_{m=1}^{M} CS_{im} / M \tag{18}$$

here $i = 1, 2, 3, 4, 5, 6$ correspond to $^{13}C_O$, $^{13}C_\alpha$, $^{13}C_\beta$, $^{1}H_N$, $^{1}H_\alpha$ and ^{15}N, respectively. M denotes the total number of residues with

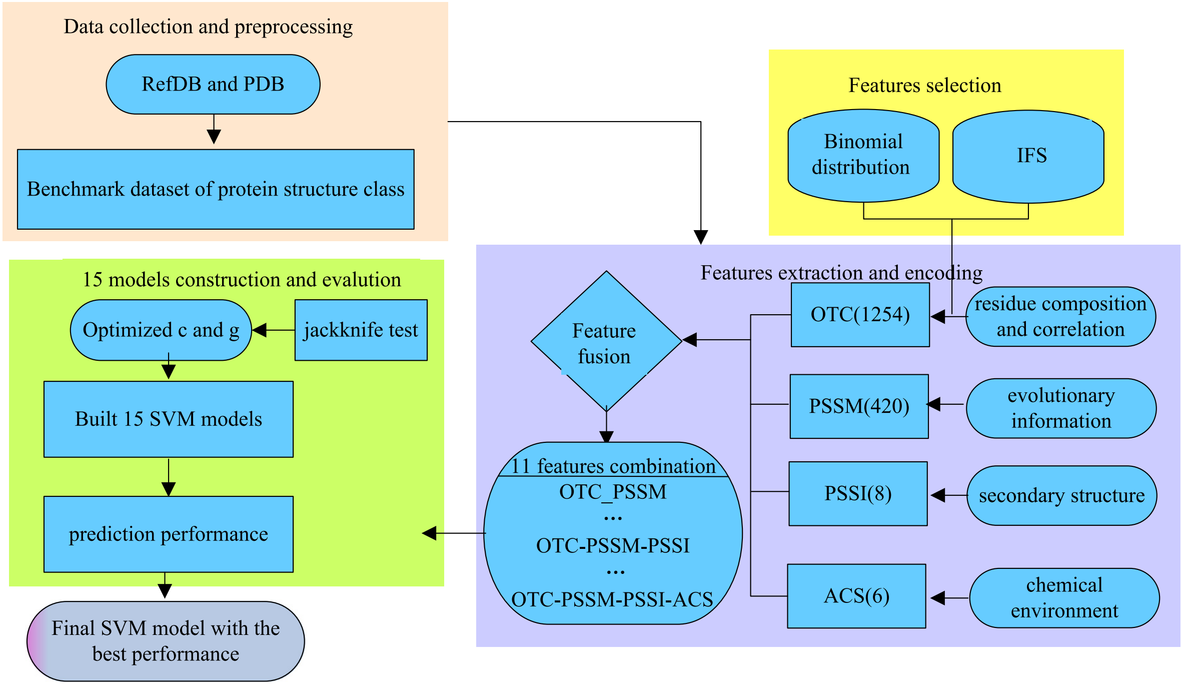

Fig. 2. The general framework for selecting optimal model.

chemical shift values assigned for nucleus species *i*. CS_{im} denotes the chemical shift value of the *i*-th nucleus at the *m*th residue. Thus, each sample can be described by a 6-dimensional feature vector expressed as:

$$\mathbf{P}_{ACS} = (ACS_1, ACS_2, \ldots, ACS_6)^T \tag{19}$$

2.6. Support vector machine

SVM is a popular machine learning technique and has been applied in protein structural class prediction [50–52] and other bioinformatics fields [53–62]. The basic idea of SVM is to project data of a sample into a high dimensional Hilbert space and to explore an optimal separating hyperplane in this space. The implementation of SVM was carried out by using the LibSVM package 3.22, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Generally, four kernel functions, namely linear function, polynomial function, sigmoid function and radial basis function (RBF), can be used in classification. Empirical studies have shown that RBF is superior to other kernel functions. Hence, we choose the RBF to perform prediction. The one-versus-one (OVO) strategy is used for multiclass classification. The regularization parameter C and kernel parameter γ were optimized through grid search with 10-fold cross validation in the LibSVM software [63]. All calculations were implemented on Ubuntu 14.04LTS running on two servers with TITAN X GPU and Intel Xeon(R) E5-2687 W v2 CPU.

2.7. Performance evaluation

In statistical prediction, three cross-validation methods are often used to examine a predictor for its effectiveness, including independent dataset set, sub-sampling test, and jackknife test [64]. In the jackknife test, one sample was selected as test dataset and the rest was regarded as training dataset. This process was repeated until all samples were examined. Due to it can yield a unique result, jackknife test has been widely used to study the accuracy of various predictors [50,51,65–70]. Hence, we adopt it to evaluate the performance of our method.

The four standard performance measures, including sensitivity (Sn), specificity (Sp), overall accuracy (OA), and average prediction accuracy (AA), were used to evaluate the performance and defined as:

$$Sn_j = \frac{TP_j}{TP_j + FN_j} \tag{20}$$

$$Sp_j = \frac{TN_j}{TN_j + FP_j} \tag{21}$$

$$OA = \sum_{j=1}^{n} \frac{TP_j}{N} \tag{22}$$

$$AA = \sum_{j=1}^{n} \frac{Sn_j}{n} \tag{23}$$

where TP_j, TN_j, FP_j, and FN_j respectively denote true positives, true negatives, false positives, false negatives of the *j*th structural class, N and n represent the number of total samples and number of structural classes, respectively.

We obtained the available and non-redundant features and constructed the high accuracy of model. An intuitive picture to describe the general framework is shown in Fig. 2.

3. Result and discussion

We firstly investigated the prediction performances of four kinds of features namely OTC, PSSM, PSSI and ACS by using SVM. For tripeptide composition optimized by binomial distribution, the IFS technique was used to obtain the OTC which could produce the maximum accuracy. The details of optimization process can be referred to Section 2.2. The IFS curve was drawn in Fig. 3. We noticed that when 1254 optimal tripeptides were used, the maximum overall accuracy was obtained. As a result, the overall accuracies are 91.0%, 70.7%, 89.2%, 86.7% for OTC, PSSM, PSSI and ACS, respectively.

Subsequently, we examined the prediction accuracies of all combinations of four features. Thus, a total of 11 experiments

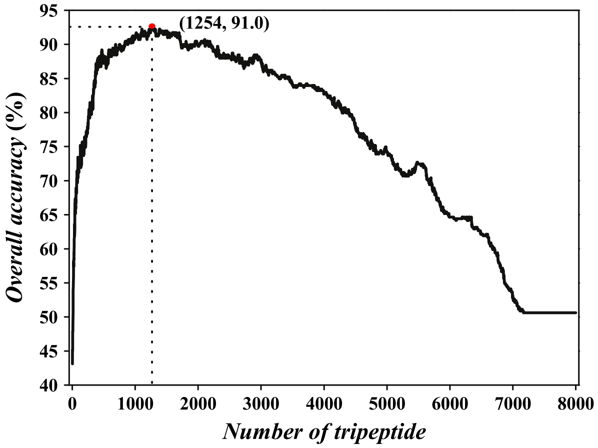

Fig. 3. A plot showing the IFS curve. When the top1254 tripeptides were used to perform prediction, the overall success rate reaches a peak of 91.0%.

Table 1
The overall accuracies based on different feature sets.

Feature	OA (%)	Feature	OA (%)
OTC	91.0	PSSM-ACS	83.2
PSSM	70.7	PSSI-ACS	93.5
PSSI	89.2	OTC-PSSM-PSSI	93.7
ACS	86.7	OTC-PSSM-ACS	95.7
OTC-PSSM	90.0	OTC-PSSI-ACS	95.5
OTC-PSSI	96.0	PSSM-PSSI-ACS	89.0
OTC-ACS	96.7	OTC-PSSM-PSSI-ACS	95.0
PSSM-PSSI	84.2		

Table 2
The prediction qualities of four best models.

Features	Structural class	Sn (%)	Sp (%)	OA (%)	AA (%)
OTC-PSSI	All-α	96.0	96.0	96.0	95.5
	All-β	91.1	97.9		
	$\alpha\beta$	99.4	93.6		
OTC-ACS	**All-α**	**96.8**	**96.7**	**96.7**	**96.4**
	All-β	**92.9**	**98.3**		
	αβ	**99.4**	**94.9**		
OTC-PSSM-ACS	All-α	96.8	95.3	95.7	95.3
	All-β	90.2	97.9		
	$\alpha\beta$	98.8	93.6		
OTC-PSSI-ACS	All-α	95.2	95.6	95.5	94.9
	All-β	90.2	97.6		
	$\alpha\beta$	99.4	92.8		
OTC-PSSM-PSSI-ACS	All-α	96.8	94.2	95.0	94.7
	All-β	91.1	96.5		
	$\alpha\beta$	96.3	94.1		

Table 3
The comparison of different method for predicting protein classes.

Method	Sensitivity (%)			Overall accuracy (%)
	All-α	All-β	$\alpha\beta$	
Dipeptides [11]	46.8	33.9	65.6	50.9
Dipeptides+AAC [11]	49.2	35.7	66.9	52.6
PseAAC [11]	68.5	59.8	65.6	64.9
Optimal tetrapeptides [39]	91.9	75.0	84.0	84.0
Optimal ACS [11]	92.7	77.7	91.4	88.0
OTC-ACS (This paper)	**96.8**	**92.9**	**99.4**	**96.7**

$(C_4^2 + C_4^3 + C_4^4 = 11)$ were performed for searching the optimal feature combination. The results were recorded in Table 1.

It was found that five types of feature combinations (OTC-PSSI, OTC-ACS, OTC-PSSM-ACS, OTC-PSSI-ACS, OTC-PSSM-PSSI-ACS) could yield the overall accuracies of >95%. Another three feature combinations (OTC-PSSM, PSSI-ACS and OTC-PSSM-PSSI) obtained the overall accuracies of >90%. It was noticed that the accuracy by using all features (OTC-PSSM-PSSI-ACS) is not the best one among all combinations. In contrast, the best prediction accuracies (96.7%) were obtained by using OTC-ACS. These results suggest that there is information redundancy or noise in the feature combination. In fact, the performance of PSSM is the worst when comparing with OTC, PSSI and ACS. Moreover, OTC, PSSI and ACS could reflect the intrinsic properties of protein structural classes in the aspect of residue composition and correlation, structure as well as chemical characteristics. Thus, it is not surprising that the OTC-PSSI and OTC-ACS based models are superior to other models. This result suggests that the evolution information is not important feature for protein structural classes.

For the five models with higher accuracies, we reported their sensitivities, specificities and average accuracies in Table 2.

It is necessary to investigate whether our proposed method has a better performance than other existing approaches. Thus, the results of published methods for the same aim were all listed in Table 3. It was found that the method proposed in this paper is superior to other published methods. Optimal tetrapeptides and optimal ACS could also achieve encouraging results. For all-α, the sensitivity of our method is about 4.9%, 4.1% higher than that of optimal tetrapeptides and optimal ACS. For all-β, the sensitivity of our method is about 17.9%, 15.2%higher than that of optimal tetrapeptides and optimal ACS. And for $\alpha\beta$ class, our method can produce the sensitivity of 99.4% which is about 15.4% and 8.0% higher than that of optimal tetrapeptides and optimal ACS.

From Tables 1 to 3, one may notice that OTC is the best feature for protein structural classes prediction. Some studies showed that the tripeptides can be utilized in discovering peptides and small organic molecule mimics [71], predicting plausible structures for oligopeptides as well as denovo protein design [72]. Thus, these could be used to explain why OTC could produce the maximum accuracy of 91.0%. PSSI is the second best feature for the prediction because the protein structural classes correlate with the content of protein secondary structure. ACS could describe the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. Thus, it has also been recognized as powerful indicators of macromolecular structure.

4. Conclusion

In this work, we investigated the accuracies of different features (i.e., residue composition and correlation, evolution, secondary structure and chemical environment) for identifying the protein structure class for low similarity sequences. Since most of the existing methods are all built by using SVM, for a fair comparison, SVM is used in the current study to perform predictions. By a deal of experiments, we found that the maximum accuracy was achieved by combining optimal tripeptide composition with chemical shift feature, indicating that evolution information is not a very important feature for protein structural class prediction. Comparative results demonstrate that the proposed method outperforms the previous published methods. Thus, the method can be used as a reliable tool for the accurate prediction of protein structural class for low-similarity sequences. Particularly, it has not escaped our notice that some computational intelligence algorithms have been developed in the past several years [73–104]. Therefore, we will develop smart models by using the newest computational intelligence algorithms to predict protein structure classes in our future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61772119, 31771471), the Fundamental Research

Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118), Natural Science Foundation for Distinguished Young Scholar of Hebei Province, China (No. C2017209244), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province, China (No. BJ2014028) and the Science Strength Promotion Programme of UESTC, China.

Competing interests

The authors declare that there are no competing interests regarding the publication of this paper.

Author contributions

H.L. conceived and designed the experiments; X.J.Z., C.Q.F., H.Y.L. and W.C. analyzed the data and implemented SVM. X.J.Z., W.C. and H.L performed the analysis and wrote the paper. All authors read and approved the final manuscript.

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–558.

[2] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH–a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.

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

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

[5] K.C. Chou, Energy-optimized structure of antifreeze protein and its binding mechanism, J. Mol. Biol. 223 (1992) 509–517.

[6] H. Cid, M. Bunster, M. Canales, F. Gazitua, Hydrophobicity and structural classes in proteins, Protein Eng. 5 (1992) 373–375.

[7] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, G. Valiente, Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment, BMC Bioinf. 8 (2007) 252.

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

[9] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. Sattar, Proposing a highly accurate protein structural class predictor using segmentation-based features, BMC Genomics 15 Suppl 1 (2014) S2.

[10] L. Kong, L. Zhang, Novel structure-driven features for accurate prediction of protein structural class, Genomics 103 (2014) 292–297.

[11] H. Lin, C. Ding, Q. Song, P. Yang, H. Ding, K.J. Deng, W. Chen, The prediction of protein structural class using averaged chemical shifts, J. Biomol. Struct. Dyn. 29 (2012) 643–649.

[12] S. Xie, Z. Li, H. Hu, Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization, Gene 642 (2018) 74–83.

[13] L. Zhang, L. Kong, X. Han, J. Lv, Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure, J. Theoret. Biol. 400 (2016) 1–10.

[14] P.P. Zhu, W.C. Li, Z.J. Zhong, E.Z. Deng, H. Ding, W. Chen, H. Lin, Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition, Mol. Biosyst. 11 (2015) 558–563.

[15] J.N. Song, K. Burrage, Predicting residue-wise contact orders in proteins by support vector regression, BMC Bioinform. 7 (2006) 425.

[16] W. Bao, Y. Chen, D. Wang, Prediction of protein structure classes with flexible neural tree, Bio-med. Mater. Eng. 24 (2014) 3797–3806.

[17] W.M. Liu, K.C. Chou, Prediction of protein structural classes by modified mahalanobis discriminant algorithm, J. Protein Chem. 17 (1998) 209–217.

[18] M.H. Olyaee, A. Yaghoubi, M. Yaghoobi, Predicting protein structural classes based on complex networks and recurrence analysis, J. Theoret. Biol. 404 (2016) 375–382.

[19] H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, J. Comput. Chem. 28 (2007) 1463–1466.

[20] Z. Aydin, A. Singh, J. Bilmes, W.S. Noble, Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure, BMC Bioinform. 12 (2011) 154.

[21] A. Chinnasamy, W.K. Sung, A. Mittal, Protein structure and fold prediction using Tree-Augmented naive Bayesian classifier, J. Bioinforma. Comput. Biol. 3 (2005) 803–819.

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

[23] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 153–162.

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

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

[26] 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.

[27] S. Costantini, A.M. Facchiano, Prediction of the protein structural class by specific peptide frequencies, Biochimie 91 (2009) 226–229.

[28] S.S. Sahu, G. Panda, A novel feature representation method based on chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.

[29] 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 functional annotations, PLoS One 9 (2014) e92863.

[30] 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.

[31] B. Liao, T. Peng, H. Chen, Y. Lin, Incorporating secondary structural features into sequence information for predicting protein structural class, Protein Pept. Lett. 20 (2013) 1079–1087.

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

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

[34] Z. Feng, X. Hu, Z. Jiang, H. Song, M.A. Ashraf, The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements, Saudi J. Biol. Sci. 23 (2016) 189–197.

[35] A.V. Kumar, R.F. Ali, Y. Cao, V.V. Krishnan, Application of data mining tools for classification of protein structural class from residue based averaged NMR chemical shifts, Biochim. Biophys. Acta 1854 (2015) 1545–1552.

[36] G. Zhou, X. Xu, C.T. Zhang, A weighting method for predicting protein structural class from amino acid composition, Eur. J. Biochem. 210 (1992) 747–749.

[37] W.S. Bu, Z.P. Feng, Z.D. Zhang, C.T. Zhang, Prediction of protein (domain) structural classes based on amino-acid index, Eur. J. Biochem. 266 (1999) 1043–1049.

[38] Y. Liang, S. Liu, S. Zhang, Prediction of protein structural classes for low-similarity sequences based on consensus sequence and segmented PSSM, Comput Math. Methods Med. 2015 (2015) 370756.

[39] H. Ding, H. Lin, W. Chen, Z.Q. Li, F.B. Guo, J. Huang, N.N. Rao, Prediction of protein structural classes based on feature selection technique, Interdiscip. Sci.-Comput. Life Sci. 6 (2014) 235–240.

[40] H. Zhang, S. Neal, D.S. Wishart, RefDB: a database of uniformly referenced protein chemical shifts, J. Biomol. NMR 25 (2003) 173–195.

[41] H.M. Berman, The protein data bank: a historical perspective, Acta Crystallogr. A 64 (2008) 88–95.

[42] G. Wang, R.L. Dunbrack, Jr, PISCES: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.

[43] Z. Chen, P. Zhao, F.Y. Li, A. Leier, T.T. Marquez-Lago, Y.N. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J.N. Song, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.

[44] F. Li, C. Li, J. Revote, Y. Zhang, G.I. Webb, J. Li, J. Song, T. Lithgow, GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features, Sci. Rep. 6 (2016) 34595.

[45] F.Y. Li, C. Li, M.J. Wang, G.I. Webb, Y. Zhang, J.C. Whisstock, J.N. Song, GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome, Bioinformatics 31 (2015) 1411–1419.

[46] 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.

[47] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.

[48] S.P. Mielke, V.V. Krishnan, Protein structural class identification directly from NMR spectra using averaged chemical shifts, Bioinformatics 19 (2003) 2054–2064.

[49] S.P. Mielke, V.V. Krishnan, Characterization of protein secondary structure from NMR chemical shifts, Prog. Nucl. Mag. Res. Sp. 54 (2009) 141–165.

[50] 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.

[51] Y. Liang, S. Zhang, Predict protein structural class by incorporating two different modes of evolutionary information into chou's general pseudo amino acid composition, J Molecul. Graphics Modell. 78 (2017) 110–117.

[52] M. Nasrul Islam, S. Iqbal, A.R. Katebi, M. Tamjidul Hoque, A balanced secondary structure predictor, J. Theoret. Biol. 389 (2016) 60–71.

[53] Y.W. Zhao, Z.D. Su, W. Yang, H. Lin, W. Chen, H. Tang, 2.0 IonchanPred : A Tool to Predict Ion Channels and Their Types, Int. J. Mol. Sci. 18 (2017) 1838.

[54] H. Lin, Z.Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE/ACM Trans. Comput. Biol. Bioinform. (2017) http://dx.doi.org/10.1109/TCBB.2017.2666141.

[55] H.Y. Lai, X.X. Chen, W. Chen, H. Tang, H. Lin, Sequence-based predictive modeling to identify cancerlectins, Oncotarget 8 (2017) 28169–28175.

[56] H. Yang, H. Tang, X.X. Chen, C.J. Zhang, P.P. Zhu, H. Ding, W. Chen, H. Lin, Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition, BioMed Res. Int. 2016 (2016) 5413903.

[57] X.X. Chen, H. Tang, W.C. Li, H. Wu, W. Chen, H. Ding, H. Lin, Identification of bacterial cell wall lyases via pseudo amino acid composition, BioMed Res. Int. 2016 (2016) 1654623.

[58] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, K.C. Chou, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. (2018) http://dx.doi.org/10.1093/bib/bby028.

[59] M.J. Wang, X.M. Zhao, H. Tan, T. Akutsu, J.C. Whisstock, J.N. Song, Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets, Bioinformatics 30 (2014) 71–80.

[60] J.N. Song, H. Tan, A.J. Perry, T. Akutsu, G.I. Webb, J.C. Whisstock, R.N. Pike, PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites, PLoS One 7 (2012) e50300.

[61] J.N. Song, H. Tan, H.B. Shen, K. Mahmood, S.E. Boyd, G.I. Webb, T. Akutsu, J.C. Whisstock, Cascleave: towards more accurate prediction of caspase substrate cleavage sites, Bioinformatics 26 (2010) 752–760.

[62] J.N. Song, H. Tan, K. Mahmood, R.H.P. Law, A.M. Buckle, G.I. Webb, T. Akutsu, J.C. Whisstock, Prodepth: predict residue depth by support vector regression approach from protein sequences only, PLoS One 4 (2009) e7072.

[63] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:21–27:27.

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

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

[66] L. Kong, L. Zhang, J. Lv, Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition, J. Theoret. Biol. 344 (2014) 12–18.

[67] L. Zhang, X. Zhao, L. Kong, A protein structural class prediction method based on novel features, Biochimie 95 (2013) 1741–1744.

[68] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (2017) 3518–3523.

[69] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iSS-PseDNC: Identifying splicing sites using pseudo dinucleotide composition, Biomed. Res. Int. 2014 (2014) ID 623149.

[70] P.-M. Feng, W. Chen, H. Lin, K.-C. Chou, iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[71] P. Ung, D.A. Winkler, Tripeptide motifs in biology: targets for peptidomimetic design, J. Med. Chem. 54 (2011) 1111–1125.

[72] S. Anishetty, G. Pennathur, R. Anishetty, Tripeptide analysis of protein structures, BMC Struct. Biol. 2 (2002) 9.

[73] R.M. Rizk-Allah, R.A. El-Sehiemy, G.G. Wang, A novel parallel hurricane optimization algorithm for secure emission/economic load dispatch solution, Appl. Soft Comput. 63 (2018) 206–222.

[74] B.Q. Li, Y.H. Zhang, M.L. Jin, T. Huang, Y.D. Cai, Prediction of protein-peptide interactions with a nearest neighbor algorithm, Curr. Bioinform. 13 (2018) 14–24.

[75] Z.H. Cui, F. Xue, X.J. Cai, Y. Cao, G.G. Wang, J.J. Chen, Detection of malicious code variants based on deep learning, IEEE Trans. Ind. Inform. 14 (2018) 3187–3196.

[76] L.Z. Yuan, E.F. Yong, Z. Wei, K.G. Shan, Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts, Curr. Bioinform. 12 (2017) 52–56.

[77] S. Patel, R. Tripathi, V. Kumari, P. Varadwaj, DeepInteract: deep neural network based protein-protein interaction prediction tool, Curr. Bioinform. 12 (2017) 551–557.

[78] I. Naseem, S. Khan, R. Togneri, M. Bennamoun, ECMSRC: a sparse learning approach for the prediction of extracellular matrix proteins, Curr. Bioinform. 12 (2017) 361–368.

[79] X.G. Nan, L.L. Bao, X.S. Zhao, X.W. Zhao, A.K. Sangaiah, G.G. Wang, Z.Q. Ma, EPuL: an enhanced positive-unlabeled learning algorithm for the prediction of pupylation sites, Molecules 22 (2017) 1463.

[80] H.X. Long, M. Wang, H.Y. Fu, Deep convolutional neural networks for predicting hydroxyproline in proteins, Curr. Bioinform. 12 (2017) 233–238.

[81] K. Liu, D.W. Gong, F.L. Meng, H.H. Chen, G.G. Wang, Gesture segmentation based on a two-phase estimation of distribution algorithm, Inform. Sci. 394 (2017) 88–105.

[82] Y.Q. Lin, X.P. Min, L.L. Li, H. Yu, S.X. Ge, J. Zhang, N.S. Xia, Using a machine-learning approach to predict discontinuous antibody-specific B-cell epitopes, Curr. Bioinform. 12 (2017) 406–415.

[83] Z.H. Cui, B. Sun, G.G. Wang, Y. Xue, J.J. Chen, A novel oriented cuckoo search algorithm to improve DV-Hop performance for cyber-physical systems, J. Parallel Distrib Com. 103 (2017) 42–52.

[84] J.H. Yi, J. Wang, G.G. Wang, Improved probabilistic neural networks with self-adaptive strategies for transformer fault diagnosis problem, Adv. Mech. Eng. 8 (2016) 1–13.

[85] G.G. Wang, A.H. Gandomi, X.J. Zhao, H.C.E. Chu, Hybridizing harmony search algorithm with cuckoo search for global numerical optimization, Soft Comput. 20 (2016) 273–285.

[86] G.G. Wang, A.H. Gandomi, X.S. Yang, A.H. Alavi, A new hybrid method based on krill herd and cuckoo search for global optimisation tasks, Int. J. Bio-Inspir. Com. 8 (2016) 286–299.

[87] H. Yang, W.R. Qiu, G.Q. Liu, F.B. Guo, W. Chen, K.C. Chou, H. Lin, iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. 14 (2018) 883–891.

[88] H. Yang, H. Lv, H. Ding, W. Chen, H. Lin, iRNA-2OM: A sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens, J. Comput. Biol. (2018) http://dx.doi.org/10.1089/cmb.2018.0004.

[89] W. Tang, S.X. Wan, Z. Yang, A.E. Teschendorff, Q. Zou, Tumor origin detection with tissue-specific miRNA and DNA methylation markers, Bioinformatics 34 (2018) 398–406.

[90] H. Tang, Y.W. Zhao, P. Zou, C.M. Zhang, R. Chen, P. Huang, H. Lin, HBPred: a tool to identify growth hormone-binding proteins, Int J Biol Sci 14 (2018) 957–964.

[91] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, K.C. Chou, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics (2018) http://dx.doi.org/10.1093/bioinformatics/bty508.

[92] B. Manavalan, T.H. Shin, G. Lee, DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest, Oncotarget 9 (2018) 1944–1956.

[93] B. Manavalan, T.H. Shin, G. Lee, PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine, Front Microbiol 9 (2018) 476.

[94] W.Y. He, C.Z. Jia, Y.C. Duan, Q. Zou, 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features, BMC. Syst. Biol. 12 (2018) 44.

[95] R.Z. Cao, B. Adhikari, D. Bhattacharya, M. Sun, J. Hou, J.L. Cheng, QAcon: single model quality assessment using protein structural and contact information with machine learning techniques, Bioinformatics 33 (2017) 586–588.

[96] R. Cao, C. Freitas, L. Chan, M. Sun, H. Jiang, Z. Chen, ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network, Molecules 22 (2017) 1732.

[97] Q. Zou, J.C. Zeng, L.J. Cao, R.R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.

[98] Q. Zou, S.X. Wan, Y. Ju, J.J. Tang, X.X. Zeng, Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, BMC Syst. Biol. 10 (2016) 114.

[99] H. Tang, W. Chen, H. Lin, Identification of immunoglobulins using chou's pseudo amino acid composition with feature selection technique, Molecular BioSyst. 12 (2016) 1269–1275.

[100] R.Z. Cao, D. Bhattacharya, J. Hou, J.L. Cheng, DeepQA: improving the estimation of single protein model quality with deep belief networks, BMC Bioinform. 17 (2016) 495.

[101] R. Cao, Z. Wang, Y. Wang, J. Cheng, SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines, BMC Bioinform. 15 (2014) 120.

[102] P.M. Feng, H. Lin, W. Chen, Identification of antioxidants from sequence information using naive Bayes, Comput. Math. Methods Med. 2013 (2013) 567529.

[103] P.M. Feng, H. Ding, W. Chen, H. Lin, Naive Bayes classifier with feature selection to identify phage virion proteins, Comput. Math. Methods Med. 2013 (2013) 530696.

[104] Y.H. Feng, G.G. Wang, Binary moth search algorithm for discounted {0-1} Knapsack Problem, IEEE Access 6 (2018) 10708–10719.