Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 1650693, 13 pages
https://doi.org/10.1155/2022/1650693

Research Article

Comparative Study on Feature Selection in Protein Structure and Function Prediction

Wenjing Yi,[1] Ao Sun,[2] Manman Liu,[2] Xiaoqing Liu,[3] Wei Zhang,[2] and Qi Dai[1]

[1]*College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[2]*College of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[3]*College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China*

Correspondence should be addressed to Wei Zhang; zhangweicse@zstu.edu.cn and Qi Dai; daiailiu04@yahoo.com

Received 27 July 2022; Accepted 14 September 2022; Published 11 October 2022

Academic Editor: Lin Lu

Many effective methods extract and fuse different protein features to study the relationship between protein sequence, structure, and function, but different methods have preferences in solving the research of protein structure and function, which requires selecting valuable and contributing features to design more effective prediction methods. This work mainly focused on the feature selection methods in the study of protein structure and function, and systematically compared and analyzed the efficiency of different feature selection methods in the prediction of protein structures, protein disorders, protein molecular chaperones, and protein solubility. The results show that the feature selection method based on nonlinear SVM performs best in protein structure prediction, protein solubility prediction, protein molecular chaperone prediction, and protein solubility prediction. After selection, the accuracy of features is improved by 13.16% ~71%, especially the Kmer features and PSSM features of proteins.

1. Introduction

Protein structure and function is the basic research field of protein research, which is of great significance for the study of protein folding rate, DNA binding sites, and protein folding recognition [1–7]. In recent years, the gap between protein sequence and protein structure is becoming larger and larger with the development of sequencing technology, and the speed of identifying protein structure and function through experimental methods is relatively slow. Therefore, it is necessary to develop computational methods to quickly and accurately determine protein structure and function.

The function of a protein is determined by its spatial structure, which is determined by its sequence. Therefore, sequence information can be used to predict protein structure and function directly, so as to further guide biological experiments and reduce experimental costs. After the concept of protein structure class was put forward, several protein structure and function prediction methods were proposed [3–5, 7–11]. Some methods use protein composition information to predict protein structure and function [1, 12, 13]. For example, short peptide composition [14–16], pseudo amino acid composition [17–20], and functional domain composition match [21]. The sequence characteristic information is expressed as amino acid composition (AAC) by calculating the ratio of 20 amino acid residues in the sequence [14–16], but it does not take into account the physicochemical properties and interaction of amino acids. In order to overcome the above problems, pseudo amino acid composition (PseACC) calculates the composition of amino acid residues based on the hydrophobicity and other physical and chemical properties of amino acid residues [17–23].

The above methods are outstanding in high similarity data, but for low similarity data, their performance is ordinary, with prediction accuracy 50%. Therefore, we need to design more effective prediction algorithm. Kurgan et al. predicted protein secondary structures and designed SCPRED method on this basis [24]. Zhang et al. calculated the TPM matrix and took it as the characteristic representation of the protein secondary structures [25]. Dai et al. statistically analyzed the characteristic distribution of protein secondary structures and applied them to protein structure prediction [26]. Ding et al. constructed a

multidimensional representation vector of protein secondary structure features, and fused it with existing features to achieve protein structure prediction [27]. Chen et al. and Kumar et al. combined structural information with physical and chemical characteristics to design a protein structure prediction method [28, 29]. Nanni et al. calculated the primary sequence characteristics and secondary structure characteristics of protein, respectively, for protein structure and function prediction [30]. Wang et al. simplified PSSM features and combined them with protein secondary structure features for protein structure prediction [31].

Through the fusion of the above features, the prediction accuracy of some methods on low similarity data sets has been improved to more than 80%, but there are still some problems in the development of protein structure and function prediction. In order to improve the prediction accuracy and efficiency of the model, the existing research is mainly achieved by fusing different types of protein features. However, it is worth noting that a simple combination of different features does not necessarily improve the prediction performance. If the combination is not appropriate, it may even offset the information contained in each other, which will not only lead to information redundancy but also increase the complexity and calculation of the model. This requires selecting valuable and contributing features, and then effective fusion, in order to design more effective prediction methods of protein structure and function.

With the above problems in mind, we introduced 16 feature selection methods based on mutual information, feature selection based on support vector machine, feature selection based on genetic algorithm, feature selection based on kurtosis and skewness, ReliefF, and sequentialfs information selection, and systematically compared their performance in protein structure class prediction, protein disorder prediction, protein molecular chaperone prediction, and protein solubility prediction. Through a comprehensive comparison and discussion, some novel valuable guidelines for use of the feature selection method for protein structure and function prediction are obtained.

2. Materials and Methods

2.1. Datasets. Four standard data sets for protein structure and function prediction were used in this work, which are protein structural class data set, molecular chaperone data set [32], solubility data set [33], and protein disorder data set [34]. The structure data set consists of 278 α structural proteins and 192 β proteins composition of structure. The molecular chaperone data set [35] is composed of 109 proteins that need Dnak/GroEL molecular chaperones to fold correctly, and 39 proteins that can fold autonomously. The solubility data set [36] is composed of 1000 proteins with high solubility and 1000 proteins with low solubility. The protein disorder data set is composed of 630 disordered proteins from DisProt and 3347 structural proteins from SCOP [37]. The detailed information of the data set is shown in Table 1.

Table 1: Detailed information of protein structure and function data.

Dataset	Positive	Negative	Total
Protein structural class	278	192	470
Molecular chaperone	109	39	148
Solubility	1000	1000	2000
Protein disorder	630	3347	3977

2.2. Sequence Feature. Six kinds of different characteristic information of proteins are extracted [26]. They are Kmer, Pseudo Amino Acid Composition (PseAAC), Correlation-based features (correlation), composition-transition-distribution (RCTD), order-based features (order), position-based features (position), GO, and position-specific score matrix (PSSM).

2.3. Feature Selection

2.3.1. Feature Selection Based on Mutual Information. Feature selection based on mutual information has become more and more popular in data mining, especially because of their ease of use, effectiveness, and strong theoretical foundation rooted in information theory. We adopted nine feature selection algorithms based on mutual information [38], which are maxRelFS, MRMRFS, minRedFS, MIQFS, QPFSFS, SPECCMI_Fs, MRMTRFS, CMIMFSand, and CIFEFS. The common point of these methods is that they all focus on the concepts of redundancy and correlation, and use greedy schemes to build the selected feature sets incrementally. Given a sample, the column is the characteristic matrix X, and the corresponding category is C. The calculation formula of mutual information is

$$Rel(X_i) = I(X_i; C) = \sum_{X_i, C} P(X_i, C) \log \frac{P(X_i, C)}{P(X_i)P(C)}. \quad (1)$$

If the selected set is S, the calculation formula of redundancy is as follows:

$$Red(X_i|S) = \frac{1}{S} \sum_{X_j \in S} I(X_i; X_j). \quad (2)$$

The above nine feature selection algorithms calculate the mutual information value of each feature and category C, and select the feature with the largest mutual information as the optimal feature. Then, according to the feature selection method of quadratic programming, the features with minimum redundancy and maximum correlation are selected one by one. Finally, we can get a feature vector sorted according to the importance of features.

2.3.2. Support Vector Machine Recursive Feature Extraction (SVM-RFE). Support vector machine recursive feature extraction (SVM-RFE) is divided into linear SVM-RFE and nonlinear SVM-RFE. The details are as follows:

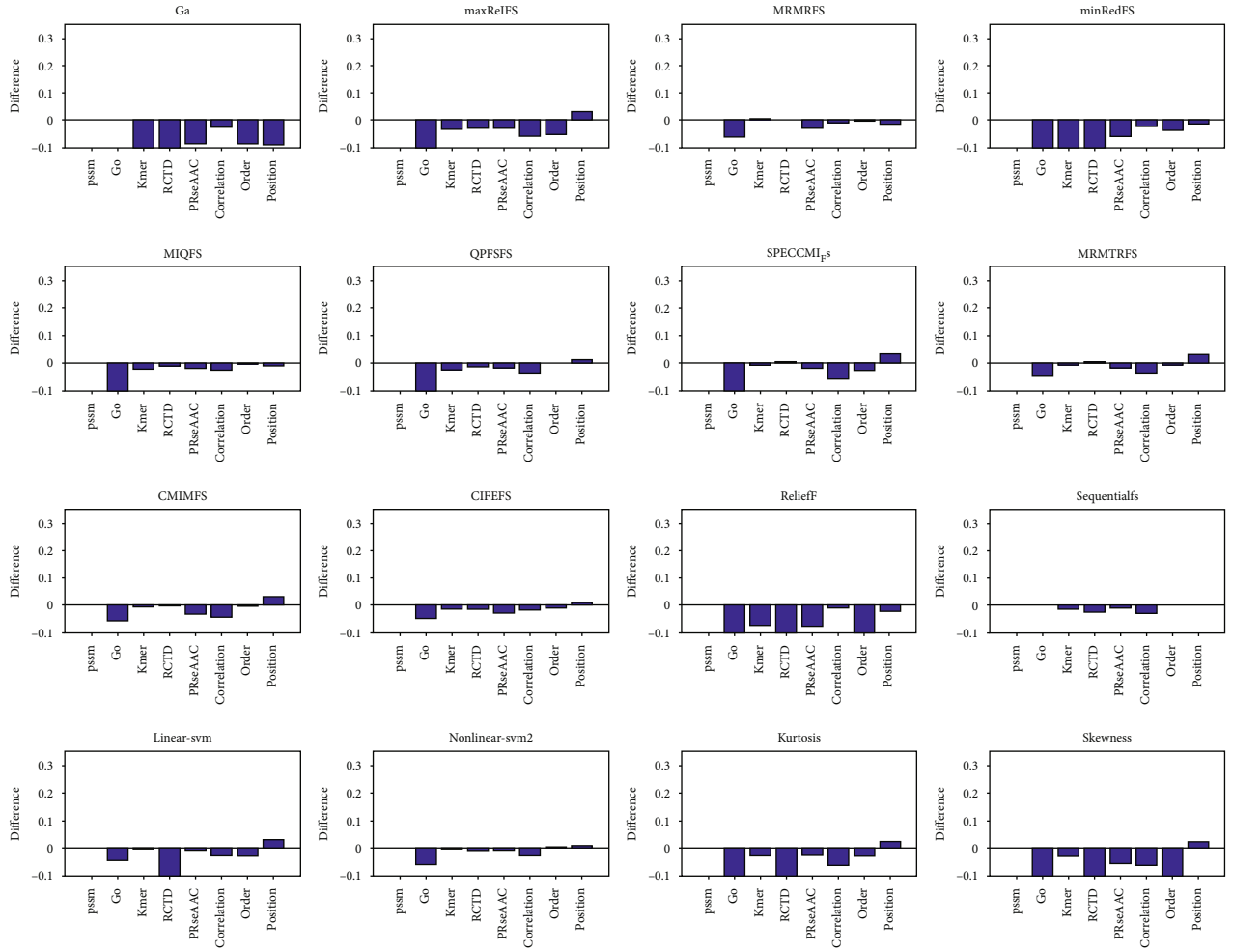

Figure 1: The comparison between the accuracy of support vector machine prediction and that of single class feature prediction after selecting the top 10 features. For each graph, the selection method is arranged from left to right and from top to bottom. They are GA, and there are nine selection methods of mutual information, relief, sequentialfs, linear SVM, nonlinear SVM, kurtosis, and skewness. The horizontal axis represents sequence features, which are PSSM, go, Kmer, RCTD, PRseAAC, correlation, order, and position, respectively.

(1) Linear SVM-RFE. For a samples {x_i, yi}, the objective function of linear SVM-RFE is

$$f(x) = a \cdot x + b, \tag{3}$$

where a is weight factor and b is deviation. Thus, the Lagrangian version of this problem can be expressed as

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j, \tag{4}$$

where α_i is Lagrange factor. α_i can be calculated by LD maximum under the condition of $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Weighting factors can be calculated by the following formula:

$$a = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{5}$$

k-th feature sorting criteria is the square of the k-th weighting factor.

$$J(k) = w_k^2. \tag{6}$$

In the training process, the feature with the smallest influence factor will be deleted every time, and so on, until all the features are deleted. Then, the importance of features is sorted according to the order in which they are deleted [39].

(2) Nonlinear SVM-RFE. In many cases, the number of features of the sample will be more than the number of samples. At this time, using linear SVM-RFE can avoid the phenomenon of over fitting [40]. However, when the number of samples is greater than the number of features, the selection result of nonlinear SVM-RFE will be better than that of linear SVM-RFE.

Nonlinear SVM-RFE will map features to new spaces with higher dimensions as follows:

$$x \in R^d \mapsto \varphi(x) \in R^h. \tag{7}$$

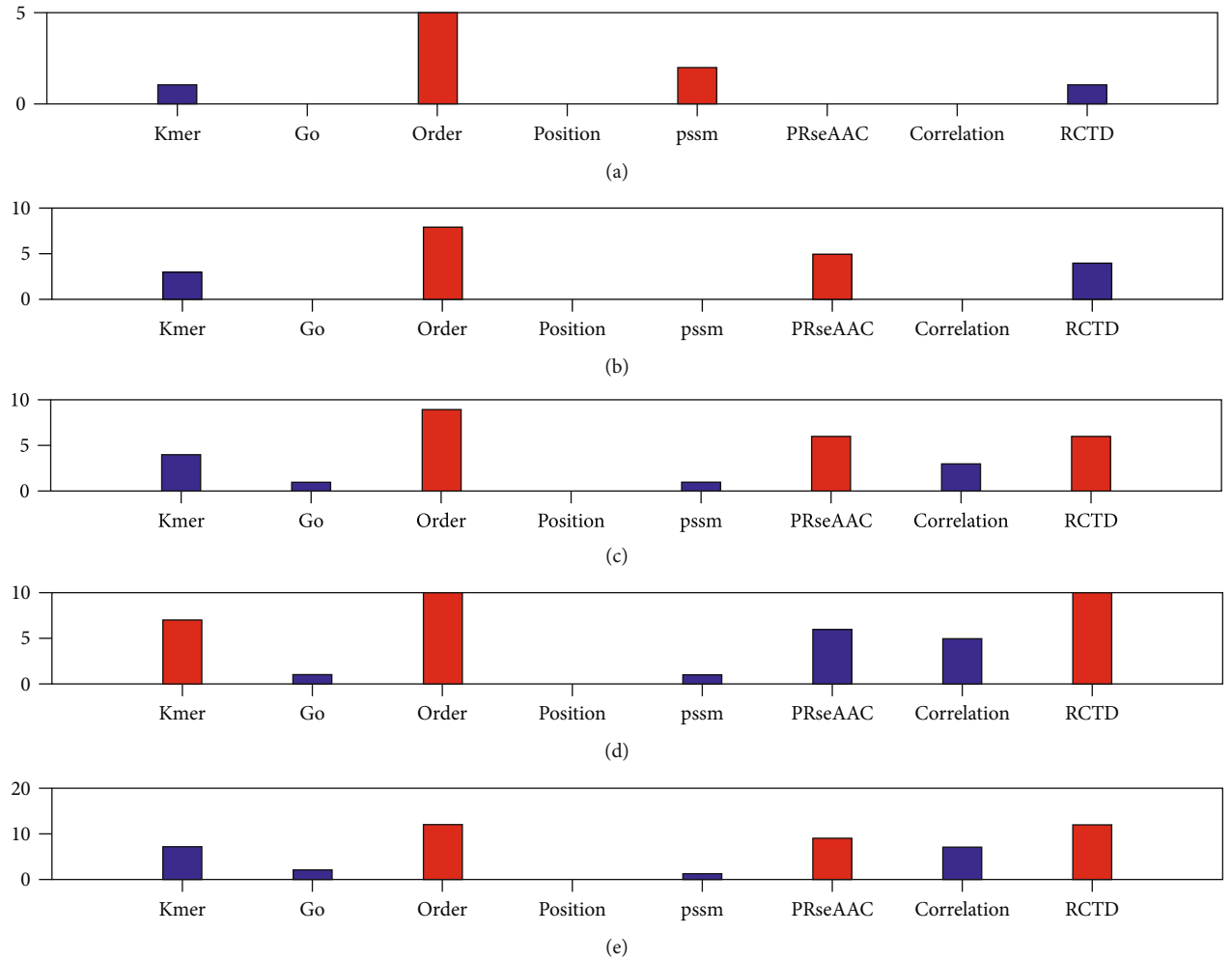

Figure 2: The number of 8 types of features in the top selected features in the protein structural data. From (a) to (e), it means that the number of selected features is 10 to 50, respectively.

In the new space, the samples are expected to be linearly separable. Its Lagrangian form can be expressed as

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i) \varphi(x_j). \tag{8}$$

Thus, we could transform inner product $\varphi(x_i)\varphi(x_j)$ into a Gaussian kernel $K(x_i, x_j)$ as follows:

$$K(x_i, x_j) = \ell^{-\lambda \left\| x_i - x_j \right\|^2}. \tag{9}$$

Thus, k-th feature sorting criteria could be expressed as

$$J(k) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K\left(x_i^{(-k)}, x_j^{(-k)}\right). \tag{10}$$

$x_i^{(-k)}$ represents that feature k has been removed.

2.3.3. Feature Selection Based on Genetic Algorithm. We adopted the assembled neural network (ASNN) algorithm. This method carries out combinatorial optimization by using the idea of genetic algorithm. For a given data set, a behavior sample can be constructed and listed as the matrix X of features [41], and finally a feature vector will be obtained, which is the optimal feature set, but the ranking of each feature is not related to its importance.

2.3.4. Feature Selection Based on Kurtosis and Skewness. For a vector of length n $\{x_1, x_2, ..., x_n\}$, its kurtosis and skewness are calculated as follows:

$$\begin{aligned} \text{Kurtosis} &= \frac{\sum_{i=1}^{n} (x_i - x)^4}{(n-1)SD^4} - 3, \\ \text{skewness} &= \frac{\sum_{i=1}^{n} (x_i - x)^3}{(n-1)SD^3}. \end{aligned} \tag{11}$$

Kurtosis and skewness are statistics used to measure the distribution of data. In this work, we calculated the skewness

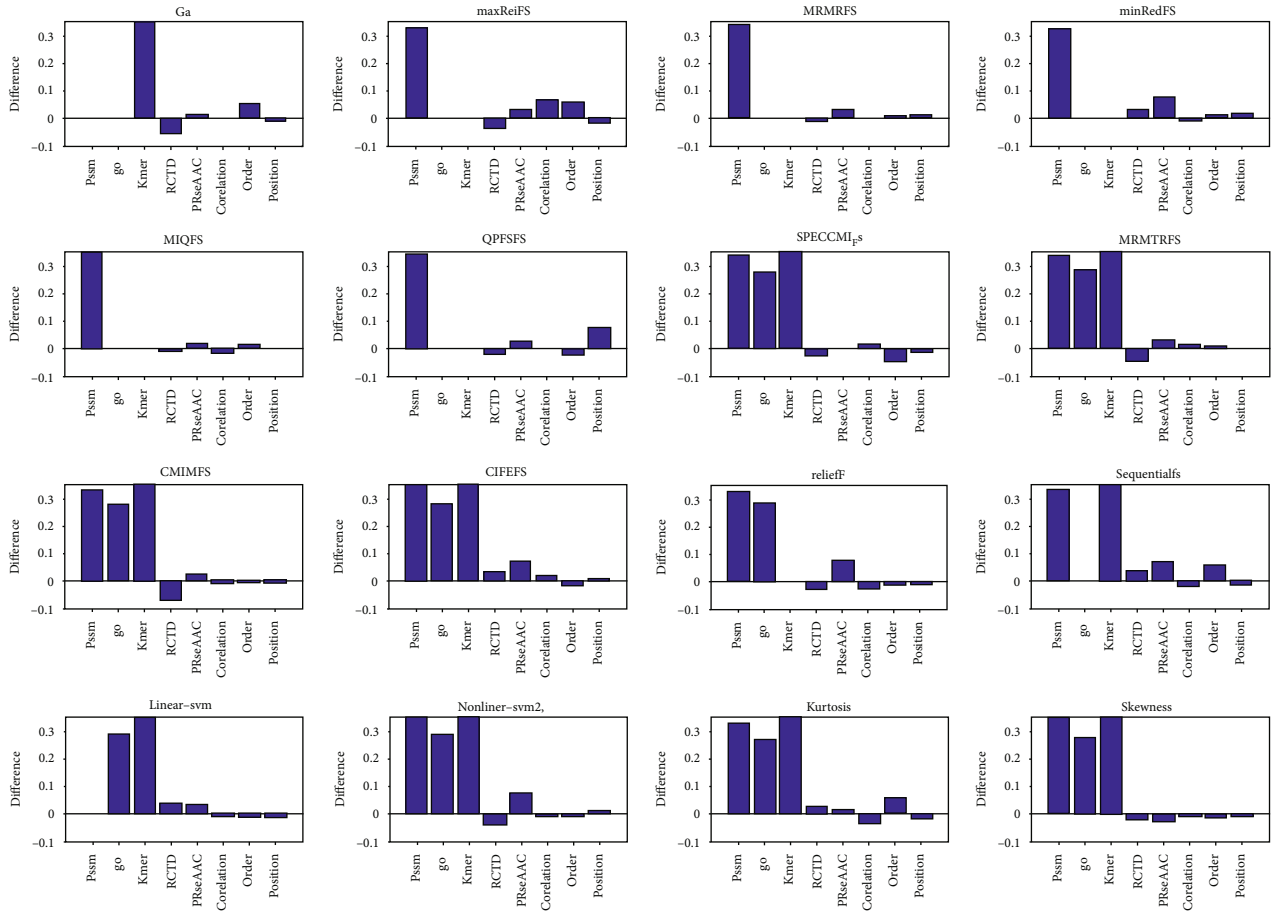

FIGURE 3: The comparison between the accuracy of support vector machine prediction and that of single class feature prediction after selecting the top 10 features. For each graph, the selection method is arranged from left to right and from top to bottom. They are GA, and there are nine selection methods of mutual information, relief, sequentialfs, linear SVM, nonlinear SVM, kurtosis and skewness. The horizontal axis represents sequence features, which are PSSM, go, Kmer, RCTD, PRseAAC, correlation, order, and position, respectively.

and kurtosis of each feature, and then sort them according to their values as a method to measure the importance of features.

2.3.5. Relieff Algorithm. Relieff algorithm randomly takes a sample R from the training sample set every time, then finds k nearest neighbor samples of R from the sample set of the same kind as R, and finds k nearest neighbor samples from the sample set of different classes of each R, and then updates the weight of each feature. The formula is as follows:

$$W(A) = W(A) - \sum_{j=1}^{k} \frac{\mathrm{diff}\left(A, R, H_j\right)}{mk} + \sum_{C \notin \mathrm{class}(R)} \frac{p(C)/1 - p(\mathrm{Class}(R)) \sum_{j=1}^{k} diff\left(A, R, M_j(C)\right)}{mk}, \tag{12}$$

where $\mathrm{diff}(A, R1, R2)$ means the difference between feature $R1$ and $R2$ in feature A. $M_j(C)$ means the j-th nearest neighbor sample in class C. Formula is as follows:

$$\mathrm{diff}(A, R1, R2) = \begin{cases} \dfrac{|R_1[A] - R_2[A]|}{\max(A) - \min(A)} & \text{if } A \text{ is consequent,} \\ 0 & \text{if } A \text{ is unconsequent and } R_1[A] = R_2[A], \\ 1 & \text{if } A \text{ is unconsequent and } R_1[A] \neq R_2[A]. \end{cases} \tag{13}$$

2.3.6. Sequentialfs. We adopted the forward feature selection algorithm of sequence in this work. For a training set $\{x_{\mathrm{train}}, y_{\mathrm{train}}\}$ and validation set $\{x_{\mathrm{validation}}, y_{\mathrm{validation}}\}$, the evaluation criteria can be expressed as

$$\left\| y_{\mathrm{validation}} - x_{\mathrm{validation}} \frac{x_{\mathrm{train}}}{y_{\mathrm{train}}} \right\|_2. \tag{14}$$

2.4. Classification Algorithm. Support vector machine is a large-scale edge classifier based on statistical learning theory [42]. It uses the optimal separation hyperplane to separate two kinds of data. For binary support vector machines, the decision function is

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b. \tag{15}$$

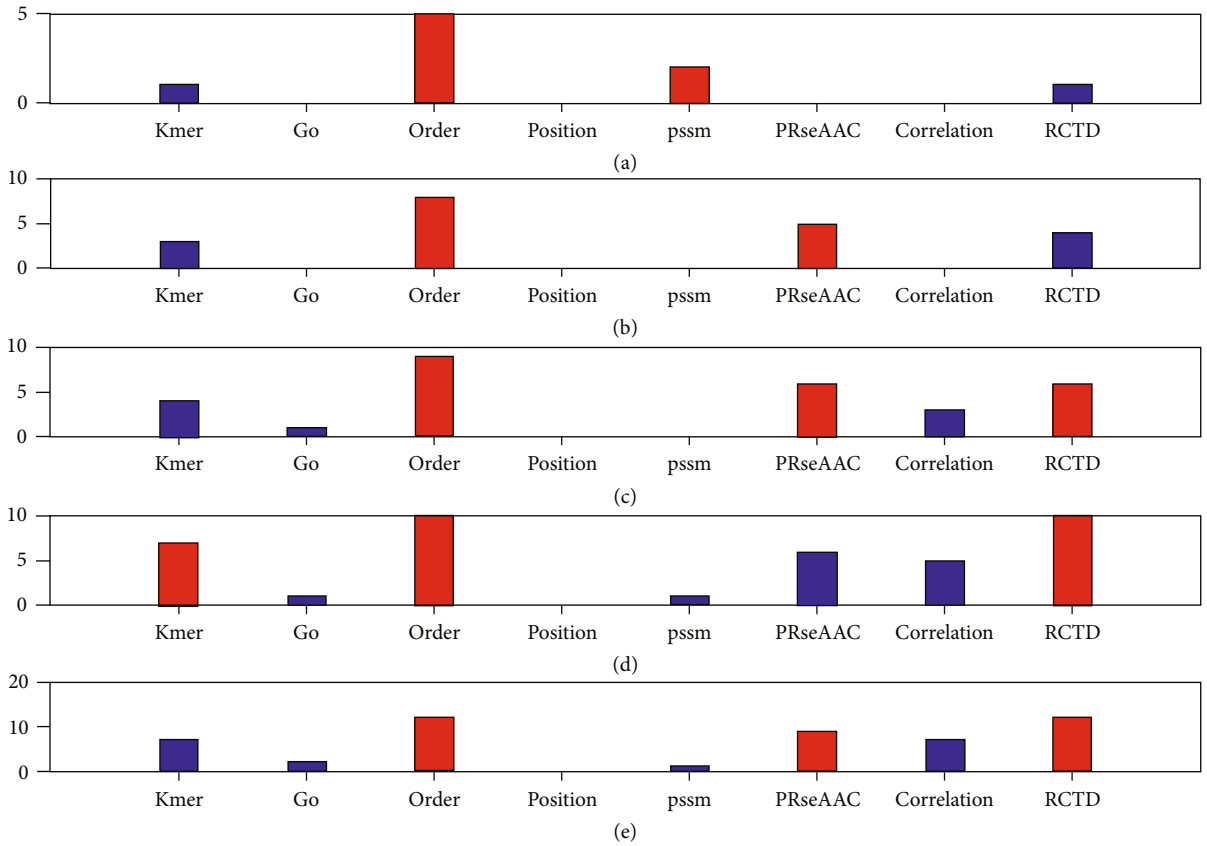

Figure 4: The number of 8 types of features in the top selected features in the protein structural data. From (a) to (e), it means that the number of selected features is 10 to 50, respectively.

where b is a constant, C is a cost parameter controlling the trade-off between allowing training errors and forcing rigid margins, $y_i \epsilon \{-1,+1\}$, x_i is the support vector, $0 \le \alpha_i \le C$, and $K(x_i, x)$ is the kernel function. This work chooses the Gauss kernel function of support vector machine because of its superiority in solving nonlinear problems [34, 37]. Furthermore, a simple grid search strategy is used to select the parameters C and gamma with the highest overall prediction. It is designed based on 10 times cross validation of each dataset, and the values of C and gamma are taken from the 2^{-10} to 2^{10}.

2.5. Performance Evaluation. This work adopted different feature selection methods for different data sets, and used the leave one method for evaluation. Finally, the prediction results are compared by calculating accuracy.

For each data set, we compared the efficiency of different feature selection methods through the following steps. The following takes the feature selection method based on genetic algorithm (GA) and PSSM features as examples to introduce the evaluation process:

(1) PSSM information is selected by GA feature selection method

(2) Select the top 10, 20, 30, 40, and 50 features using GA (if the number of features is insufficient, all the information will be taken out), input them into SVM classifier for classification prediction, and calculate the accuracy of prediction ACC1, ACC2, ACC3, ACC4, and ACC5

(3) Subtract the accuracy of the whole PSSM information from ACC1, ACC2, ACC 3, ACC 4, and ACC 5

(4) Compare the changes in accuracy of various special products after 16 selection methods

We also compared and analyzed the characteristics of selection, and the main steps are as follows:

(1) Use the above 16 selection methods to select each type of feature

(2) According to the selection results of 16 feature selection methods, the importance of each type of feature is ranked

(3) Take out the first 10, 20, 30, 40, and 50 features of each type of feature, respectively, (if the number of features is insufficient, all the information will be taken out) and mix them together as five new mixed features ($I_{10}, I_{20}, I_{30}, I_{40}, I_{50}$);

(4) Then, 16 feature selection methods are used to select the mixed features

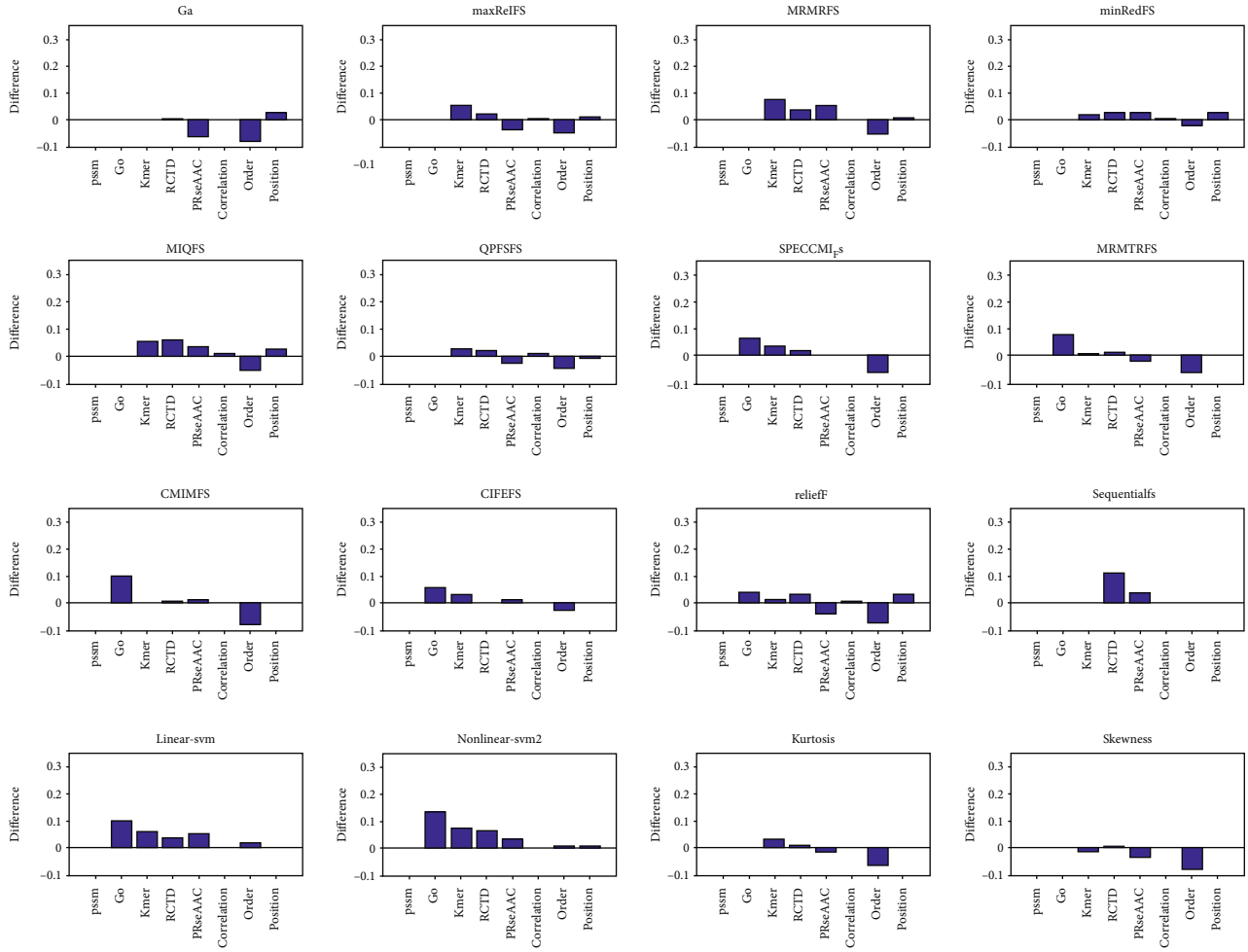

Figure 5: The comparison between the accuracy of support vector machine prediction and that of single class feature prediction after selecting the top 10 features. For each graph, the selection method is arranged from left to right and from top to bottom. They are GA, and there are nine selection methods of mutual information, relief, sequentialfs, linear SVM, nonlinear SVM, kurtosis, and skewness. The horizontal axis represents sequence features, which are PSSM, go, Kmer, RCTD, PRseAAC, correlation, order, and position, respectively.

(5) According to the results of the fourth step, the importance of the fused features is ranked

(6) Take out the top 10 features and count the type of features from which these 10 features come. Take out the top 20 features and count the categories of features

(7) In five cases, if there are a large number of certain features (or observed), it means that such features are more important

3. Results and Discussion

3.1. Comparison of Feature Selection in Protein Structure Prediction. We first discussed the efficiency of different feature selection methods in protein structure prediction. We adopted the structural data set, which contains 278 items α structural proteins and 192 β structural proteins. In this work, eight kinds of features are selected through 16 feature selection methods, and the selected features are input into the support vector machine to predict the structural class of protein. The quality of feature selection methods is evaluated based on the accuracy of prediction, which are represented in Figure 1 and Supplementary Figures 1–4.

From Figure 1 and Supplementary Figures 1–4, it is easy to note that the accuracy of MRMRFS, MRMTRFS, CMIMFS, CIFEFS, and nonlinear SVM feature selection methods changes the most, and the change range is 3.19% for the position feature. By comparing the accuracy of the first 20-50 features selected with that of the unselected features, it can be seen that the biggest change in accuracy is the GO features selected by nonlinear SVM, with changes of 2.13%, 6.39%, 6.17%, and 4.68%, respectively. Therefore, nonlinear SVM feature selection method performs best in protein structure prediction.

For structural data sets [43], we further compared and analyzed the types of selected features. First, eight types of features are fused, and the fused features are selected through 16 feature selection methods, and the top 10-50 features are

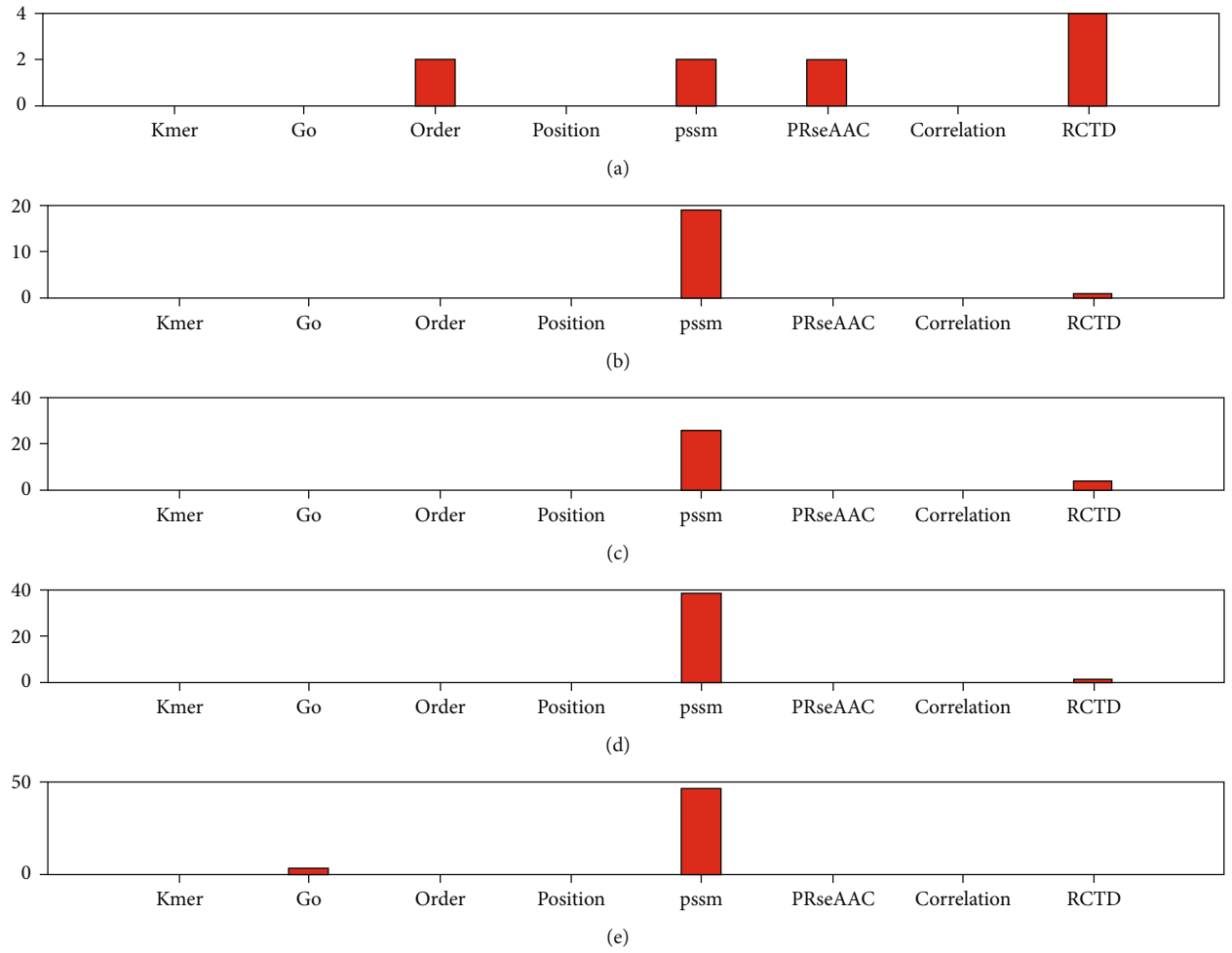

Figure 6: The number of 8 types of features in the top selected features in the protein structural data. From (a) to (e), it means that the number of selected features is 10 to 50, respectively.

selected. Then, the number of eight types of features in the top 10-50 total selected features is counted, and the preference of eight types of features is evaluated by proportion. Figure 2 shows the number of 8 types of features in the top 10-50 total selected features in the protein structural data.

Figure 2 show that when the total number of selections is 10, there are 5 order features, accounting for 50%. When the total selection number is 20, there are 8 order features, accounting for 40%. When the selection number is 40, both order and RCTD have 10, accounting for 25% of the top 40 features. When the total selection number is 50, there are 12 orders and RCTD features, respectively, accounting for 24% of the total. The above results show that order feature is the first choice for protein structure prediction, followed by RCTD feature.

3.2. Comparison of Feature Selection in Protein Disorder Prediction. We then discussed the efficiency of different feature selection methods in protein disorder prediction. The protein disorder data set [44] used in this chapter is from two protein databases related to structural classes, including 630 disordered proteins from disProt and 3347 structural proteins from SCOP. In this work, eight kinds of features are selected through 16 feature selection methods, and the selected features are input into KNN to predict protein disorder. The quality of feature selection methods is evaluated based on the accuracy of prediction, which are represented in Figure 3 and Supplementary Figures 5–8.

It can be seen from Figure 3 and Supplementary Figures 5–8 that when PSSM feature, go feature and Kmer feature are input into KNN algorithm for prediction, the change values of their accuracy are 51.28%, 55.11% and 26.95%, respectively. It can be seen that after feature selection, the accuracy of protein disorder prediction is significantly improved. When selecting 10 features, SPECCMI_FS performs best based on Kmer feature, and its accuracy by 71%. When selecting the first 20 and 30 features, the nonlinear SVM feature selection method is particularly prominent in Kmer features, and its accuracy has increased by 64.19%. Among the top 40 features selected, CIFEFS selection method performs best in Kmer features, and the accuracy is improved to 65.21%. Among the top 50 features selected, CIFEFS and linear SVM selection methods are outstanding, and the accuracy has increased by 59.61%. The above results show that for protein disorder data sets, SPECCMI_FS, CIFEFS, nonlinear SVM, and linear SVM

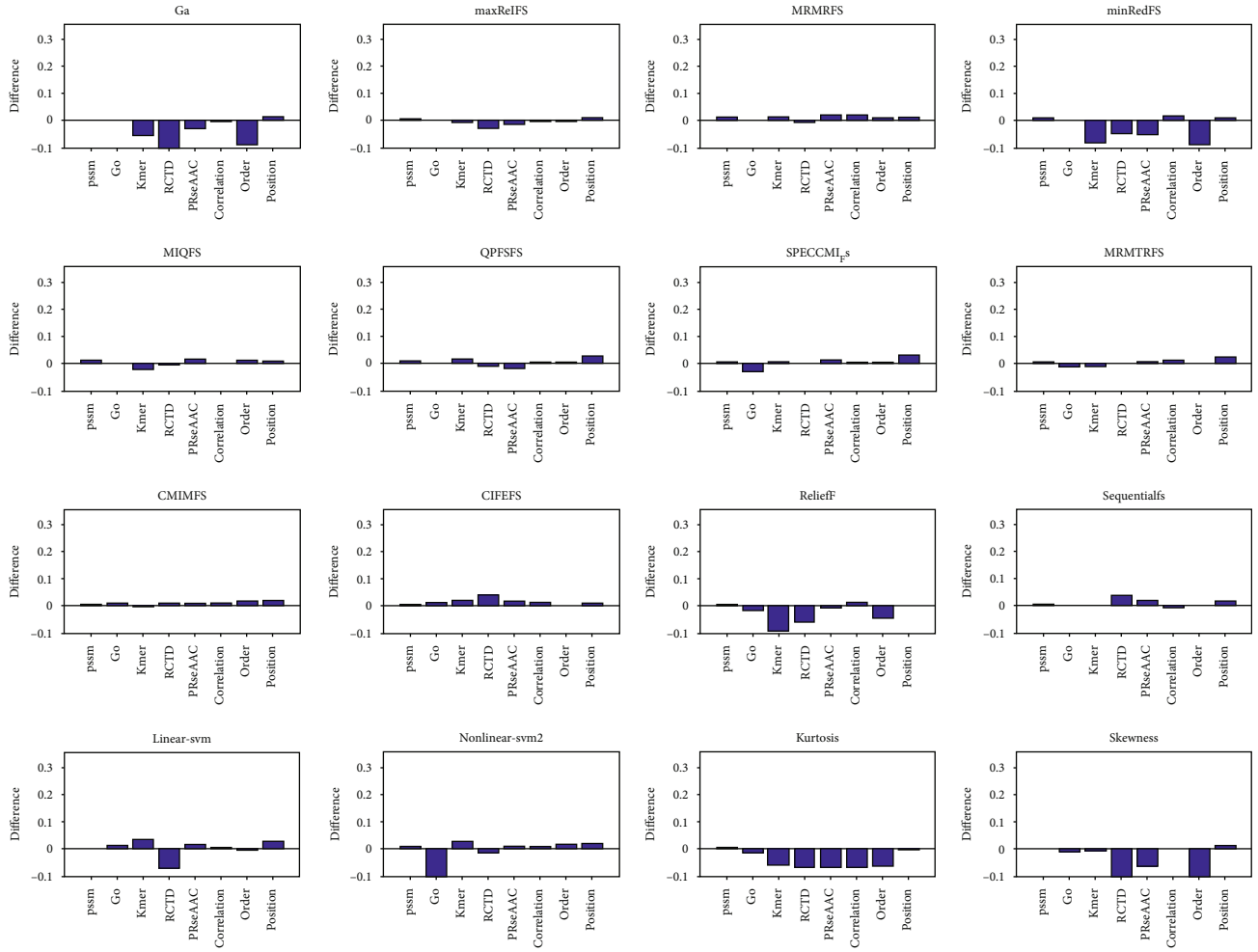

Figure 7: The comparison between the accuracy of support vector machine prediction and that of single class feature prediction after selecting the top 10 features. For each graph, the selection method is arranged from left to right and from top to bottom. They are GA, and there are nine selection methods of mutual information, relief, sequentialfs, linear SVM, nonlinear SVM, kurtosis, and skewness. The horizontal axis represents sequence features, which are PSSM, go, Kmer, RCTD, PRseAAC, correlation, order, and position, respectively.

feature selection methods can select core features from Kmer features, which improve its accuracy by 59.61% ~71%.

We also further compared the types of selected features. First, eight types of features are fused, and the fused features are selected through 16 feature selection methods, and the top 10-50 features are selected. Then, the number of eight types of features in the top 10-50 total selected features is counted, and the preference of eight types of features is evaluated by proportion. Figure 4 shows the number of 8 types of features in the top 10-50 total selected features in the protein disorder data set.

Figure 4 shows the number of features selected at five levels from top to bottom. If the top 10 fusion features are selected, 5 of them are from order features. If the first 20 fusion features are selected, 8 of them are from order features. If the first 30 fusion features are selected, 9 of them are from order features. If the first 40 fusion features are selected, there are 10 features from order and RCTD, respectively. If the top 50 fusion features are selected, 12 of them are from order and RCTD features, respectively. Therefore, the order and RCTD feature will help to improve the accuracy of the protein disorder prediction.

3.3. Comparison of Feature Selection in Protein Molecular Chaperone Prediction. We then discussed the efficiency of different feature selection methods in protein molecular chaperone prediction. In the data set used in this work, there are 109 proteins that need Dnak/GroEL molecular chaperones to fold correctly, and the remaining 39 proteins that can fold autonomously. In this work, eight kinds of features are selected through 16 feature selection methods, and the selected features are input into KNN to predict protein disorder. The quality of feature selection methods is evaluated based on the accuracy of prediction, which are represented in Figure 5 and Supplementary Figures 9–12.

Figure 5 and Supplementary Figures 9–12 show that when selecting the top 10 and 20 features, the accuracy of GO feature selection using nonlinear SVM is improved by 13.16% and 14.48%. When selecting the first 30 and 50 features, the accuracy of using sequentialfs to select RCTD

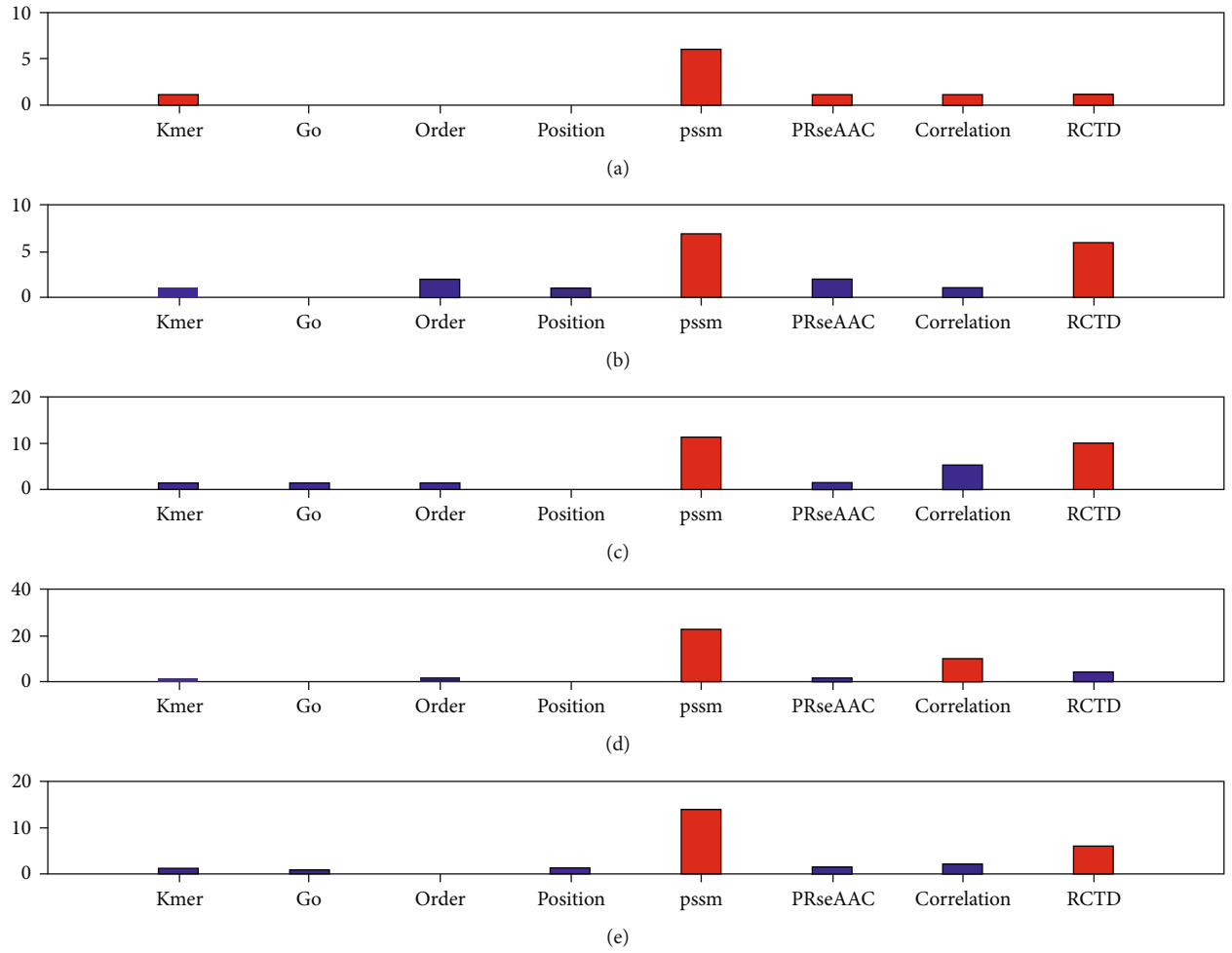

Figure 8: The number of 8 types of features in the top selected features in the protein structural data. From (a) to (e), it means that the number of selected features is 10 to 50, respectively.

Table 2: Time consumption of feature selection methods.

	Mutual Information (/S)	Sequentialfs (/S)	Linear-svm (/S)	Nonlinear-svm (/S)
PSSM	5.8	14074	23.8	2082.4
Go	360.33	—	42.4	0.75
RCTD	4.2	5571	11.7	1.3
Kmer	6.3	7423	18.9	1.7
PRseAAC	0.67	5.83	4.32	0.36
Order	1	35.2	22.8	270.5
Position	0.75	2.09	2.04	0.17
Correlation	0.62	3.87	1.39	0.33

features is improved by 13.16% and 17.17%. When selecting the first 40 features, linear SVM is used to select Kmer features, which improves its accuracy by 14.48%. Therefore, nonlinear SVM, sequentialfs and linear SVM are used to select features in the molecular chaperone prediction, which improves its accuracy by 13.16% ~17.17%.

We also further compared the types of selected features. First, eight types of features are fused, and the fused features are selected through 16 feature selection methods, and the top 10-50 features are selected. Then, the number of eight types of features in the top 10-50 total selected features is counted, and the preference of eight types of features is evaluated by proportion. Figure 6 shows the number of 8 types of features in the top 10-50 total selected features in the protein disorder data set.

When selecting 10 comprehensive features, there are 5 RCTD features, accounting for 50%. When selecting 20-30 comprehensive features, PSSM features have an absolute advantage, with 19, 26, 39, and 47 selected, respectively. It can be seen that PSSM is the preferred feature if you want to check whether a protein sequence is self-folding or molecular chaperone to help complete the correct folding.

3.4. Comparison of Feature Selection in Protein Solubility Prediction. Finally, the efficiency of different feature selection methods in protein solubility prediction is discussed. In this work, more than 7000 proteins from E. coli were selected and sorted according to their solubility. The first

1000 protein sequences with higher solubility and the last 1000 protein sequences with the lowest solubility were taken out to form a protein sequence data set. Through 16 feature selection methods, 8 kinds of features are selected, respectively, and the selected features are input into KNN to predict the solubility of protein. The quality of feature selection methods is evaluated based on the accuracy of prediction, which are represented in Figure 7 and Supplementary Figures 13–16.

When selecting 10 and 20 features, using CIFEFS based on mutual information to select RCTD features, the accuracy is improved the most, which is 3.93% and 3.88%, respectively. When selecting 30 features, using sequentialfs to select RCTD features, the accuracy is improved by 3.12%. When 40 and 50 features are selected, the accuracy of nonlinear SVM is improved by 3.12% and 4.76%, respectively. The above results show that CIFEFS, sequentialfs and nonlinear SVM feature selection methods perform well in protein solubility prediction.

We also further compared the types of selected features. First, eight types of features are fused, and the fused features are selected through 16 feature selection methods, and the top 10-50 features are selected. Then, the number of eight types of features in the top 10-50 total selected features is counted, and the preference of eight types of features is evaluated by proportion. Figure 8 shows the number of 8 types of features in the top 10-50 total selected features in the protein disorder data set.

When selecting 10-50 comprehensive features, PSSM features always account for the most, with 6, 7, 11, 23 and 28 PSSM features, accounting for 60%, 35%, 36.67%, 50.75% and 56% of the total. Therefore, using PSSM characteristics as input features to predict the solubility of new protein sequences is more reliable [45].

3.5. Comparison of Calculation Efficiency of Various Methods. The above analysis shows that the nonlinear SVM feature selection method based on support vector machine performs well in the prediction of various protein structures and functions. In order to further study the computational efficiency of feature selection methods, we calculated the time-consuming of various feature selection methods to select 8 types of features, as shown in Table 2. Mutual information represents the average time of the nine selection methods. It is not difficult to find that the nonlinear SVM selection method is related to the size of matrix elements. The larger the data elements, the longer the time required. Therefore, the matrix is normalized before feature selection. Sequentialfs consumes the most time, and the time-consuming ratio of nonlinear SVM, linear SVM, and single mutual information selection method is 2.5: 27.5 : 1. Therefore, the nonlinear SVM selection method is the preferred feature selection method in the prediction of protein structure and function.

4. Conclusion

Feature selection can reduce the problem of over fitting, improves the performance of the model, and reduces the time and space cost of the learning algorithm. 16 feature selection methods used in this work are feature selection method based on mutual information, feature selection method based on support vector machine, feature selection method based on genetic algorithm, feature selection method based on kurtosis and skewness, ReliefF ,and sequentialfs information selection methods. Different feature selection methods were compared and analyzed in protein structure class prediction, protein disorder prediction, protein molecular chaperone prediction, and protein solubility prediction.

Through a comprehensive comparison and discussion, we found that nonlinear SVM feature selection method performs best in protein structure prediction, the first choice is order feature, followed by RCTD feature. In protein disorder prediction, SPECCMI_FS, CIFEFS, nonlinear SVM, and linear SVM feature selection methods can select core features from Kmer features, which improves its accuracy by 59.61% ~71%. At the same time, order or RCTD features as input information will help to improve the accuracy of prediction. In protein molecular chaperone prediction, nonlinear SVM, sequentialfs, and linear SVM are used to select features, which improves the accuracy by 13.16% ~17.17%, and the preferred feature is PSSM feature. In protein solubility prediction, CIFEFS, sequentialfs, and nonlinear SVM feature selection methods perform well, and PSSM is the preferred feature. These results can be regarded as some novel valuable guidelines for use of the feature selection method for protein structure and function prediction.

Data Availability

The data are available in https://github.com/bioinfo0706/RaaMLab.

Conflicts of Interest

The authors declare that there are no conflicts of interest, financial, or otherwise.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62172369); Key Research and Development Plan of Zhejiang Province (2021C02039); and Zhejiang Provincial Natural Science Foundation of China (LY20F020016). The authors thank all the anonymous referees for their valuable suggestions and support.

Supplementary Materials

Supplementary Figures 1-16 are the precision comparison between support vector machine prediction and single class feature prediction based on the selected 20, 30, 40 and 50 features. *(Supplementary Materials)*

References

[1] P. Klein and C. Delisi, "Prediction of protein structural class from the amino acid sequence," *Biopolymers*, vol. 25, no. 9, pp. 1659–1672, 1986.

[2] K. C. Chou, "Structural bioinformatics and its impact to biomedical science and drug discovery," *Frontiers in Medicinal Chemistry*, vol. 3, no. 1, pp. 455–502, 2006.

[3] C. Chothia and M. Levitt, "Structural patterns in globular proteins," *Nature*, vol. 261, no. 5561, pp. 552–558, 1976.

[4] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 226D–2269, 2004.

[5] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[6] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente, "Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment," *BMC Bioinformatics*, vol. 8, no. 1, p. 252, 2007.

[7] D. Qi and T. Wang, "Comparison study on k-word statistical measures for protein: from sequence to 'sequence space'," *BMC Bioinformatics*, vol. 9, no. 1, pp. 394–394, 2008.

[8] C. Chen, Y. X. Tian, X. Y. Zou, P. X. Cai, and J. Y. Mo, "Using pseudo-amino acid composition and support vector machine to predict protein structural class," *Journal of Theoretical Biology*, vol. 243, no. 3, pp. 444–448, 2006.

[9] K. C. Chou, "Prediction of protein structural classes and subcellular locations," *Current Protein & Peptide Science*, vol. 1, no. 2, pp. 171–208, 2000.

[10] K. D. Kedarisetti, L. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology," *Biochemical and Biophysical Research Communications*, vol. 348, no. 3, pp. 981–988, 2006.

[11] D. Qi, W. Li, and L. Li, "Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features," *Journal of Computational Chemistry*, vol. 32, no. 16, pp. 3393–3398, 2011.

[12] K. C. Chou, "A key driving force in determination of protein structural classes," *Biochemical & Biophysical Research Communications*, vol. 264, no. 1, pp. 216–224, 1999.

[13] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[14] R. Y. Luo, Z. P. Feng, and J. K. Liu, "Prediction of protein structural class by amino acid and polypeptide composition," *European Journal of Biochemistry*, vol. 269, no. 17, pp. 4219–4225, 2002.

[15] X. D. Sun and R. B. Huang, "Prediction of protein structural classes using support vector machines," *Amino Acids*, vol. 30, no. 4, pp. 469–475, 2006.

[16] S. Zhang, Y. Liang, and X. Yuan, "Improving the prediction accuracy of protein structural class: approached with alternating word frequency and normalized Lempel-Ziv complexity," *Journal of Theoretical Biology*, vol. 341, no. 1, pp. 71–77, 2014.

[17] Y. S. Ding, T. L. Zhang, and K. C. Chou, "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network," *Protein and Peptide Letters*, vol. 14, no. 8, pp. 811–815, 2007.

[18] L. Wu, Q. Dai, B. Han, L. Zhu, and L. Li, "Combining sequence information and predicted secondary structural feature to predict protein structural classes," in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1–4, Wuhan, China, 2011 May 10.

[19] B. Liao, Q. Xiang, and D. Li, "Incorporating secondary features into the general form of Chou's PseAAC for predicting protein structural class," *Protein & Peptide Letters*, vol. 19, no. 11, pp. 1133–1138, 2012.

[20] L. Kong, L. Zhang, and 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," *Journal of Theoretical Biology*, vol. 344, no. 1, pp. 12–18, 2014.

[21] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 452, pp. 22–34, 2018.

[22] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, 2017.

[23] K. C. Chou and Y. D. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor," *Biochemical and Biophysical Research Communications*, vol. 320, no. 4, pp. 1236–1239, 2004.

[24] L. Kurgan, K. Cios, and K. Chen, "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinformatics*, vol. 9, no. 1, p. 226, 2008.

[25] S. Zhang, S. Ding, and T. Wang, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure," *Biochimie*, vol. 93, no. 4, pp. 710–714, 2011.

[26] Q. Dai, Y. Li, X. Liu, Y. Yao, Y. Cao, and P. He, "Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–4, 2013.

[27] H. Ding, H. Lin, W. Chen et al., "Prediction of protein structural classes based on feature selection technique," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 235–240, 2014.

[28] C. Chen, L. X. Chen, X. Y. Zou, and P. X. Cai, "Predicting protein structural class based on multi-features fusion," *Journal of Theoretical Biology*, vol. 253, no. 2, pp. 388–392, 2008.

[29] A. V. Kumar, R. F. M. Ali, Y. Cao, and V. V. Krishnan, "Application of data mining tools for classification of protein structural class from residue based averaged NMR chemical shifts," *Biochimica et Biophysica Acta*, vol. 1854, no. 10, pp. 1545–1552, 2015.

[30] L. Nanni, S. Brahnam, and A. Lumini, "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 360, pp. 109–116, 2014.

[31] J. Wang, C. Wang, J. Cao, X. Liu, Y. Yao, and Q. Dai, "Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features," *Gene*, vol. 554, no. 2, pp. 241–248, 2015.

[32] I. Antes, S. W. I. Siu, and T. Lengauer, "DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations," *Bioinformatics*, vol. 22, no. 14, pp. e16–e24, 2006.

[33] P. Klus, B. Bolognesi, F. Agostini, D. Marchese, A. Zanzoni, and G. G. Tartaglia, "The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities," *Bioinformatics*, vol. 30, no. 11, pp. 1601–1608, 2014.

[34] G. Colonna, M. Costantini, and S. Costantini, "Frequencies of specific peptides in intrinsic disordered protein domains," *Protein & Peptide Letters*, vol. 17, no. 11, pp. 1398–1402, 2010.

[35] B. Boeckmann, A. Bairoch, R. Apweiler et al., "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.

[36] C. C. H. Chang, J. Song, B. T. Tey, and R. N. Ramanan, "Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction," *Briefings in Bioinformatics*, vol. 15, no. 6, pp. 953–962, 2014.

[37] S. Idicula-Thomas and P. V. Balaji, "Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation," *Protein Engineering Design & Selection*, vol. 18, no. 4, pp. 175–180, 2005.

[38] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection," in *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, pp. 512–521, New York, New York, USA, 2014.

[39] Y. Yang and L. Chen, "Identification of drug-disease associations by using multiple drug and disease networks," *Current Bioinformatics*, vol. 17, no. 1, pp. 48–59, 2022.

[40] X. Li, L. Lu, and L. Chen, "Identification of protein functions in mouse with a label space partition method," *Mathematical Biosciences and Engineering*, vol. 19, no. 4, pp. 3820–3824, 2021.

[41] X. Pan, L. Chen, Liu, Z. Niu, T. Huang, and Y. D. Cai, "Identifying protein subcellular locations with embeddings-based node2loc," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 1–675, 2021.

[42] J. P. Zhou, L. Chen, and Z. H. Guo, "iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[43] J. Van Durme, S. Maurer-Stroh, R. Gallardo, H. Wilkinson, F. Rousseau, and J. Schymkowitz, "Accurate prediction of DnaK-peptide binding via homology modelling and experimental data," *PLoS Computational Biology*, vol. 5, no. 8, article e1000475, 2009.

[44] A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nature Biotechnology*, vol. 22, no. 10, pp. 1302–1306, 2004.

[45] J. Winkelmann, G. Calloni, S. Campioni, B. Mannini, N. Taddei, and F. Chiti, "Low-level expression of a folding-incompetent protein in *Escherichia coli*: search for the molecular determinants of protein aggregation *in vivo*," *Journal of Molecular Biology*, vol. 398, no. 4, pp. 600–613, 2010.