Genomics 112 (2020) 4666–4674

Contents lists available at ScienceDirect

Genomics

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

Prediction of antioxidant proteins using hybrid feature representation method and random forest

Chunyan Ao[a,b,1], Wenyang Zhou[c,1], Lin Gao[a], Benzhi Dong[d,*], Liang Yu[a,*]

[a] School of Computer Science and Technology, Xidian University, Xi'an, China
[b] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China
[c] Center for Bioinformatics, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China
[d] College of Computer Science and Engineering, Northeast Forestry University, Harbin, China

ARTICLE INFO

Keywords:
Antioxidant protein
Hybrid feature representation methods
MRMD
Random forest

ABSTRACT

Natural antioxidant proteins are mainly found in plants and animals, which interact to eliminate excessive free radicals and protect cells and DNA from damage, prevent and treat some diseases. Therefore, accurate identification of antioxidant proteins is important for the development of new drugs and research of related diseases. This article proposes novel method based on the combination of random forest and hybrid features that can accurately predict antioxidant proteins. Four single feature extraction methods (188D, profile-based Auto-cross covariance (ACC-PSSM), N-gram, and g-gap) and hybrid feature representation methods were used to feature extraction. Three feature selection methods (MRMD, t-SNE, and the optimal feature set selection) were adopted to determine the optimal features. The new hybrid feature vectors derived by combining 188D with the other three features all have indicators ranging from 0.9550 to 0.9990. The novel method showed better performance compared with the other methods.

1. Introduction

Natural antioxidant proteins are mainly extracted and purified from animal and plant species, such as soybean, corn, yam, potato, *Ginkgo biloba* seeds, and jellyfish [1–5], and include vitamin C, vitamin A, vitamin E, glutathione peroxidase, superoxide dismutase, and glutathione reductase [6]. Antioxidant proteins interact with free radicals (from metabolism in the body or production in response to environmental influences) to provide electrons to protect cells and DNA from damage caused by free radical-induced chain reactions, thereby prevent aging and disease [7–9]. Due to the antioxidant effects of these antioxidant proteins, they have been used to prevent aging as well as prevent and treat diseases, such as cancer, neurodegenerative diseases, heart disease, coronary heart disease and other diseases [10–17].

Obtaining antioxidant proteins through traditional biochemical methods is expensive and time-consuming [18–20]; thus, it is necessary to use machine learning to accurately identify antioxidant proteins. A machine learning method was developed to identify antioxidant proteins using star graph topological indices obtained from the protein sequence. This model was trained by a combination of the star map topological indices and random forest. The five features were extracted from the embedded map and had good predictive ability, with an accuracy of 94%. However, according to the data analysis, the protein sequence homology reached 100%, and no data redundancy was performed, which led to high prediction results [9]. The study of antioxidant protein prediction laid the foundation for subsequent research. The amino acid composition has been regarded as a critical factor in proteins that has been widely used in protein prediction [21–28]. The AAC model was used to represent the standardized frequencies of natural amino acids in the peptide chain [21]. Naïve Bayes (NB), amino acid composition and dipeptide composition feature extraction have been used to predict antioxidant proteins. The model was developed and evaluated to employ the jackknife test [29]. Pseudo-amino acid composition (PseAAC) are an improved feature extraction method on AAC, which is better than AAC for protein prediction [30,31]. Antioxidant protein prediction was performed based on the combination of a feedforward neural network and PesAAC features to evaluate the prediction model using ten-fold cross-validation [32]. G-gap dipeptide composition can be used as a feature descriptor for classification problems. This feature extraction method is based on protein sequences, which represent the long-distance correlation between two residues. A method for identifying antioxidant proteins, called AodPred is

* Corresponding authors.
E-mail addresses: nefu_dbz@163.com (B. Dong), lyu@xidian.edu.cn (L. Yu).
[1] equally contributed.

https://doi.org/10.1016/j.ygeno.2020.08.016
Received 23 July 2020; Received in revised form 10 August 2020; Accepted 13 August 2020
Available online 17 August 2020
0888-7543/

developed by combing g-gap and support vector machine (SVM) [33]. Another method using g-gap features combined with deep learning to identify antioxidant proteins and evaluate the prediction model performance using ten-fold cross-validation. IDAod uses g-gap features to identify antioxidant proteins. This method is based on ten-fold cross-validation and deep learning to select antioxidant proteins [34]. The primary sequence features (188D) were extracted from protein sequences, which are described by physicochemical properties and sequence information of the protein. Antioxidant proteins can be accurately predicted using 188D features and SVMs. The SeqSVM model uses MRMD for dimension reduction to obtain an optimal feature set, which improves the model prediction performance [35]. The accuracy of this prediction method was 89.46%. Based on evolutionary information and secondary structure information, the 473D feature extract algorithm method was proposed [36]. The 473D features were obtained through the PSI-BLAST and PSI-PRED profiles. The Aops-Svm model uses machine learning to accurately predict antioxidant proteins, which is based on sequence features and SVMs. The research indicated that the 473D feature extraction method is beneficial for the prediction of antioxidant proteins, and the accuracy rate is as high as 94.2% [37]. All of the above models used a single feature extraction method to predict antioxidant proteins. The performance of antioxidant protein prediction is related to protein expression, and different features have different effects on the performance of protein prediction. A method for predicting antioxidant proteins is proposed by g-gap and PSSM features and a mixture of the two features based on ten-fold cross-validation and random forest [38]. Another method uses hybrid feature extraction methods to predict antioxidant proteins. The feature extraction methods include Secondary Structure Information (SSI); Position Specific Scoring Matrix (PSSM); Relative Solvent Accessibility (RSA); and Composition, Transition, and Distribution (CDT). These features are incorporated together and classified using an ensemble classifier composed of RF, SMO, NNA, and J48. The results show that the performance of the model can be improved by using hybrid features and ensemble classifiers to predict antioxidant proteins [39].

To demonstrate that hybrid features can be good predictors of antioxidant proteins, a new method for predicting antioxidant proteins is proposed, which uses an extraction method of random forest combined with hybrid features. Thus, some feature extraction methods were selected, including 188D, profile-based autocross covariance (ACC-PSSM), N-gram, and g-gap. These four individual features are combined in different combination methods through the hybrid feature representation methods. Because the mixed features have redundant features, the feature selection method MRMD was used to reduce the dimensions to improve the prediction performance. The optimal feature set selection method was used to obtain the optimal features. Random forest was used as a classifier to classify antioxidant proteins and non-antioxidant proteins.

2. Methods

In this paper, the main sequence of the methods is shown in Fig. 1, which can be summarized by the following four steps.

(1) The SVMProt 188-dimensional features (188D), profile-based autocross covariance (ACC-PSSM), g-gap and N-gram ($N = 3$) were used to extract features.
(2) The four individual features are fused by the hybrid feature representation methods to obtain the following new feature sets: 188D + ACC-PSSM, 188D + N-gram, 188D + g-gap, ACC-PSSM + N-gram, ACC-PSSM + g-gap, N-gram + g-gap, 188D + ACC-PSSM + N-gram, 188D + ACC-PSSM + g-gap, 188D + N-gram + g-gap, ACC-PSSM + N-gram + g-gap, and 188D + ACC-PSSM + g-gap + N-gram.
(3) A variety of feature selection methods, including the MRMD, t-SNE and optimal feature set selection methods, are employed to obtain the optimal features through dimensionality reduction. To reduce redundant feature vectors, the MRMD method is adopted to reduce the feature dimensions. t-SNE reduced the dimensions of high-dimensional features and then visualized these features in a two-dimensional space. The optimal feature set selection method reduced the dimensions to obtain the expressive feature.
(4) The prediction model accurately predicted antioxidant proteins by combining random forest (RF) with eight-fold cross-validation.

In this paper, all classification experiments were carried out in version 3.9 of Weka. The results show that our prediction method performed better than other methods. The proposed method can accurately identify antioxidant proteins and help to study related diseases.

2.1. Benchmark dataset

In this section, we introduce the construction of benchmark datasets. Antioxidant protein sequences were retrieved from the UniProt database. The non-antioxidant protein dataset was downloaded from the research of Feng et al. [33] When we collected the antioxidant protein dataset, the following steps were taken. First, the antioxidant dataset is obtained through experimental verification. Second, non-standard letters containing “B”, “J”, “O”, “X”, or “Z” are removed from the antioxidant protein. Therefore, we obtained 546 experimentally verified antioxidant proteins as positive samples. The 1552 non-antioxidant proteins are used as negative samples [33].

If the predictors are trained and tested based on a biased baseline dataset, the predictors may produce misleading results, and their accuracy can be overestimated [40]. To avoid homology bias and reduce redundancy, the CD-Hit program [41] was used to process this raw dataset. Finally, 1984 protein sequences (containing 434 antioxidant proteins and 1550 non-antioxidant proteins) were obtained. To overcome the imbalance of the number of samples, we randomly divided the negative sample into three parts and randomly extracted 434 sequences from each part of the sample. The experimental result is the average of three experiments using these three sets.

2.2. Feature extraction

Encoding a protein sequence requires an appropriate set of information parameters, and it is one of the most important parts of identifying protein properties [42]. In this work, four individual feature extraction methods, including SVMProt 188D, Profile-based Auto-cross covariance (ACC-PSSM), N-gram, and g-gap, were used for extracting features of antioxidant protein sequences. These feature methods were based on sequence information, physiochemical properties and the correlation. The feature sizes of 188D, ACC-PSSM, N-gram and g-gap were 188, 800, 8420 and 400, respectively.

2.2.1. SVMProt 188D features

Two proteins may be structurally similar in some cases but without significant sequence similarity. Therefore, a sequence-based prediction method can accurately identify a protein [43–46]. The sequence is transformed into a feature space vector representation. In this intermediate step, different methods will affect the prediction result. According to the amino acid composition and physicochemical properties, protein characteristics were extracted from the primary sequence. Support vector machine model has been widely used in bioinformatics research [47–49]. Cai et al. [50] and Zou et al. [51] proposed using the SVMProt 188D feature extraction method. Each protein sequence extracted by this method is represented by a specific feature vector. These feature vectors are composed of the encoded representation of the attributes of the listed residues, including 20 amino acid compositions and eight physicochemical properties.

The first 20-dimensional features represent the frequency of the

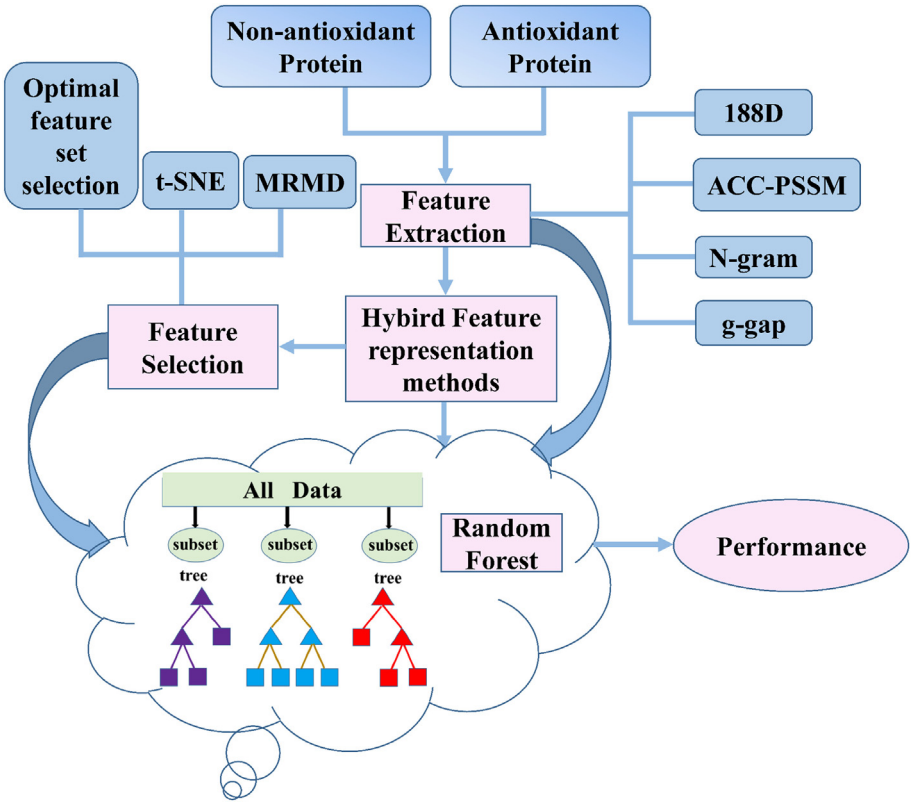

Fig. 1. Overall framework of the prediction of antioxidant proteins. First, the protein sequences are represented by 188D, ACC-PSSM, N-gram, and g-gap. Then, four feature methods are combined by the hybrid feature represented methods. Finally, a variety of feature selection methods, MRMD, t-SNE and optimal feature set selection methods, are employed to obtain optimal features through dimensionality reduction. The features obtained in the above steps are classified using random forest and eight-fold cross-validation.

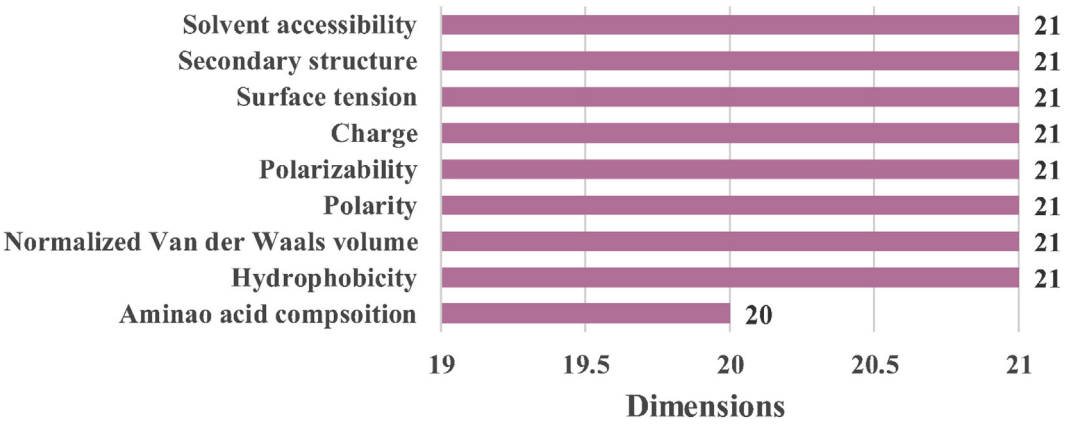

Fig. 2. 188D Feature Composition.

occurrence of 20 amino acids (alphabetically ACDEFGHIKLMNPQRSTVWY). Each property of amino acids is classified into three types according to eight physicochemical properties (including charge, surface tension, solvent accessibility, normalized Van der Waals volume, polarity, hydrophobicity, polarizability, and secondary structure). Eight kinds of physicochemical properties yielded 168-dimensional features. Therefore, we obtained 188 (20 + 168)-dimensional features using the SVMProt 188D feature extraction method. The 188D function is shown in Fig. 2.

2.2.2. Profile-based auto-cross covariance (ACC-PSSM)

ACC-PSSM is an important feature extraction method, which has been widely used in protein prediction and other bioinformatics fields [44,52–55]. Liu et al. [56] proposed the ACC-PSSM feature extraction method, which was implemented by Pse-in-one 2.0. Due to the different lengths of protein sequences, there were vectors of unequal lengths. Therefore, autocross covariance (ACC) transforms these numeric vectors into a unified matrix. ACC consists of auto covariance (AC) and cross-covariance (CC) [57]. The AC variable is used to calculate the correlation of the same properties, and *len* is denoted as the distance between two amino acid residues, which is in the same property. The AC calculation formula is as follows [58,59]:

$$\mathrm{AC}(\mathrm{x}, len) = \frac{1}{S - len} \sum_{t=1}^{S-len} \left(\mathrm{F}_{\mathrm{x},t} - \frac{1}{S} \sum_{t=1}^{S} F_{\mathrm{x},t} \right) \left(F_{\mathrm{x},t+len} - \frac{1}{S} \sum_{t=1}^{S} F_{\mathrm{x},t} \right) \quad (1)$$

where S represents the protein sequence length, F represents the protein sequence, x represents one of the twenty amino acids, and $F_{x,\ t}$ represents the score of x amino acid in PSSM at the t position. After calculation by AC, PSSM is converted into a fixed-length vector of $20{*}M_{len}$, where M_{len} is the maximum value of *len*.

The CC calculated the correlation between two different properties, for which the calculation formula is as follows [57]:

$$CC(x1, x2, len) = \frac{1}{S - len} \sum_{t=1}^{S-len} \left(F_{x1,t} - \frac{1}{s} \sum_{t=1}^{S} F_{x1,t} \right) \left(F_{x2,t+len} - \frac{1}{S} \sum_{t=1}^{S} F_{x2,t} \right) \quad (2)$$

where x1, x2 represents the x1, x2 amino acids. After calculation by CC, PSSM can be converted into a fixed-length vector of $380*M_{len}$.

2.2.3. N-gram feature

N-gram [60] is a type of probabilistic model based on Markov hypothesis. When training the model, the maximum likelihood was used to estimate the model parameter-conditional probability. When using machine learning for protein identification, the N-gram model used unaligned encoding techniques to extract features from sequences [61]. In the N-gram model, the value range of N is 1, 2, and 3. This represents the unary model, the binary model, and the ternary model, respectively. In this paper, we took N as 3 and used N-gram to calculate the probability of 3-g occurrence in the sequence. For example, suppose the protein sequence is "VILTFILTF," the values of VIL, ILT, LTF, TFI, and FIL are 1, 2, 2, 1, and 1, respectively. The other three amino acid combination values are 0.

2.2.4. g-gap

In the protein sequence, two amino acid residues separated by g intervals in the primary structure may be adjacent in three dimensional space [62]. Therefore, Lin et al. [33] proposed the g-gap dipeptide combination feature extraction method, which is based on protein sequence information, to extract features. The g-gap dipeptide composition describes long distance correlations between two residues. For example, the protein sequence is "ACDEA," and the dipeptide combination is AC, CD, DE, EA when $g = 0$, which represents the correlation between adjacent amino acid residues in the protein sequence. The dipeptide combination is AD, CE, DA when $g = 1$, which represents the correlation between the two amino acid residues separated by an amino acid residue. The dipeptide combination is AE, CA when $g = 2$, which represents the correlation between the two amino acid residues separated by two amino acid residues. In this paper, we set g to 2 for feature extraction and obtained 400-dimensional features.

2.3. Classifier

To identify whether a protein sequence is an antioxidant protein, the classification is treated as a binary classification problem. We chose common classifiers for classification: random forest (RF), Naïve Bayes (NB), and J48.

RF [63] is a machine learning method for constructing Bagging integration based on multiple decision trees. During training, the decision tree further introduces random attribute selection. RF, as an integrated classifier, has good classification ability and has been widely used in regression, classification, prediction [64,65] and other problems [66–70]. For classification problems, each decision tree is classified. By voting, the class with the most votes is the output of the entire random forest. However, the average of all decision trees is the final output of the random forest on the regression problem [71].

2.4. Feature selection

In the feature extraction section, we extracted four features of 188D, N-gram, g-gap, and ACC-PSSM. Then, these new feature vectors are obtained by fusing the four features through a hybrid feature representation. These new feature vectors or extracted features may contain redundant and noisy features, which can lead to poor performance in predicting antioxidant proteins. To improve the performance of the prediction method, the feature selection method is implemented by removing redundant and noisy features and reducing feature dimensions, thus saving computing time. In this paper, we selected three different feature selection methods, MRMD, t-SNE, and the optimal feature set selection method, to find the optimal feature.

MRMD [72] is a dimension reduction method that can balance the accuracy and stability of feature classification and prediction tasks. The MRMD dimension reduction is achieved by using Pearson's correlation coefficient [73] and Euclidean distance [74]. t-SNE [75], a dimensionality reduction technique, is suitable for analysis and visualization. t-SNE is widely used [34,76–79] in the field of biological information. In protein prediction, t-SNE is used to convert the feature vector of a protein sequence to a point in the low-dimensional space. It can be seen from the visualization that adjacent points in low-dimensional space are in similar states. The optimal feature set selection method combined with machine learning was proposed by Feng et al. [80] This method uses covariance analysis, MRMD dimensionality reduction processing, and finally principal component analysis. The extracted features are processed by this method to obtain the best feature set with less dimensionality and stronger expression.

2.5. Performance evaluation

Cross-validation is an analytical method based on classical statistics [81]. Cross-validation is a commonly used evaluation method in protein prediction [82–90] and other fields [47,91–96]. In this article, eight-fold cross-validation was used to evaluate the performance of predictive antioxidant protein methods.

There are many indicators to evaluate the accuracy of the prediction method, and we used the following indicators to evaluate the model: accuracy (ACC), F-measure, sensitivity (SN), and specificity (SP). These indicators are calculated as follows:

$$SN = \frac{TP}{TP + FN} \quad (3)$$

$$SP = \frac{TN}{TN + FP} \quad (4)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall = \frac{TP}{TP + FN} \quad (7)$$

$$F - measure = \frac{(\alpha^2 + 1) precision * recall}{\alpha^2 (precision + recall)} \quad (8)$$

TP indicates the accurate prediction of antioxidant protein samples, FP indicates the prediction of non-antioxidant proteins as antioxidant proteins, TN indicates the accurate prediction as non-antioxidant proteins, FN indicates the prediction of antioxidant proteins as non-antioxidant proteins, SN is the true positive rate, SP is the true negative rate, ACC is the classification accuracy of the classifier, and the F-measure value is defined by the harmonic mean of *precision and recall* [97]. In general, a is set to 1.

3. Results and discussion

3.1. Performance of different features and classifiers

In this section, we used four single feature representation methods, including 188D, ACC-PSSM, N-gram and g-gap, to classify protein sequences. In addition, we compare different classifiers in each single feature representation method. After screening, we determined that RF is the best method for the classification of antioxidant proteins. The results of different feature extraction methods and the results of different classifiers are shown as follows (Table 1).

According to the experimental results in Table 1, when the

Table 1
Comparison of classification results of different feature extraction methods and different classifiers.

Method	Classifier	AUC	ACC	F-Measure	SN	SP
188D	**RF**	**0.9990**	**0.9819**	**0.9820**	**0.9854**	**0.9785**
	Naïve bayes	0.9513	0.9228	0.9227	0.9478	0.8978
	J48	0.9847	0.9766	0.9767	0.9779	0.9840
ACC-PSSM	**RF**	**0.8977**	**0.8322**	**0.8313**	**0.7650**	**0.8993**
	Naïve bayes	0.7510	0.6863	0.6843	0.6130	0.7597
	J48	0.7110	0.7039	0.7057	0.7197	0.6880
N-gram(N = 3)	**RF**	**0.8930**	**0.8191**	**0.8153**	**0.8057**	**0.8327**
	Naïve bayes	0.7237	0.7135	0.7070	0.5630	0.8717
	J48	0.6907	0.6786	0.6783	0.3577	0.6783
g-gap	**RF**	**0.8890**	**0.8168**	**0.8170**	**0.8127**	**0.8210**
	Naïve bayes	0.7820	0.7012	0.7003	0.7447	0.6577
	J48	0.6923	0.6843	0.6807	0.6950	0.6660

Bold letters represent the maximum value in each comparison.

classifiers are RF, NB, and J48, the values of AUC, ACC, F-measure, sensitivity and specificity obtained by 188D feature extraction method are the highest, followed by ACC-PSSM, N-gram, and g-gap. 188D achieves the best performance, with an AUC value of 0.9990, ACC of 0.9819, F-measure of 0.9820, SN of 0.9854 and SP of 0.9785 using RF as a classifier, which comprehensively outperforms ACC-PSSM by more than 10.13% in AUC value, 14.97% in ACC, 15.07% in F-measure, 22.04% in SN, and 7.92% in SP. When RF is used as a classifier, various indicators of 188D are significantly higher than N-gram. The AUC value, ACC, F-measure, SN, SP are 10.6%, 16.28% 16.67% 17.97% 14.58% higher than N-gram, respectively. Similarly, 188D performed better than g-gap using RF as a classifier. 188D exceeds g-gap in AUC value by 11%, ACC by 16.51%, F-measure by 16.5%, SN by 17.27%, and SP by 15.75%. 188D is clearly superior to other feature extraction methods (Table 1).

According to comparison and analysis, RF shows better performance than the other two classifiers among the four feature extraction methods. The AUC value obtained by RF range from 0.9990 to 0.8890, with the highest AUC value obtained by NB of 0.9513, the lowest of 0.7273, and the AUC value obtained by J48 of between 0.9847 and 0.6907. When using different feature extraction methods, the F-measure value obtained by RF as a classifier is also the highest. In summary, RF is the most effective classifier for predicting antioxidant proteins (Table 1).

In addition to comparing the three classifiers, we also used three classifiers to vote for the combined classifiers. The Vote was executed in Weka. The classifier is composed of RF, NB, and J48 and performs majority voting. The Voting criteria were as follows: if two are correct and one is wrong, the final output is correct. The classification results are shown in Fig. 3A. With the g-gap feature extraction method, the accuracy obtained by the vote classifier is 0.7861, and the accuracies obtained by the three separate classifiers RF, NB, and J48 are 0.8168, 0.7012, and 0.6843, respectively. Although the accuracy of the vote classifier is lower than RF, it is 8.49% and 10.18% higher than NB and J48. The vote classifier was compared to the classification results of three classifiers, where RF also have the highest accuracy, and the other two classifiers have lower accuracy. Therefore, the vote classifier has the effect of correcting the weak classifier.

3.2. Performance of the hybrid feature representation methods

The four features are combined by the hybrid feature extraction method to obtain new feature vectors of different combinations. The results show that these new feature vectors have better performance compared with the single feature representation methods (Table 2). Seven new feature vectors were obtained by 188D combined with the other three features. All indicators of these seven new feature vectors are significantly higher than the other three single features (ACC-PSSM, N-gram, g-gap), which are all above 0.9000 (Table 1 **and** Table 2). When the three features of ACC-PSSM, N-gram and g-gap are combined with each other, the indicators of the new feature vector fluctuate. The hybrid feature vector may contain redundant features, which have an impact on the classification effect.

3.3. Performance of feature selection methods

Because hybrid features may include redundant features and noise, MRMD was selected for feature selection. MRMD mainly reduces the dimensionality of hybrid features, and the results are shown in Fig. 3B. The AUC of the mixed feature vectors are compared with the AUC of the mixed feature vectors after dimensionality reduction. After the hybrid features reduced the dimensionality using MRMD, the performance of predicting antioxidant proteins is improved, and the AUC value is increased. Only the AUC value of ACC-PSSM + N-gram features are reduced, which may be related to the feature extraction method.

In addition to using MRMD for feature selection, we also used the optimal feature selection method and t-SNE. The implementation of feature selection has the following two steps: first, the optimal feature is obtained by using the optimal feature selection method; second, the obtained optimal feature is reduced by t-SNE and visualized in the two-dimensional feature space. We used the optimal feature set method and t-SNE for feature selection for the four single features.

The obtained original 188D features are processed by covariance analysis, MRMD, and principal component analysis. The optimal feature set selection method reduces the feature dimension to the best

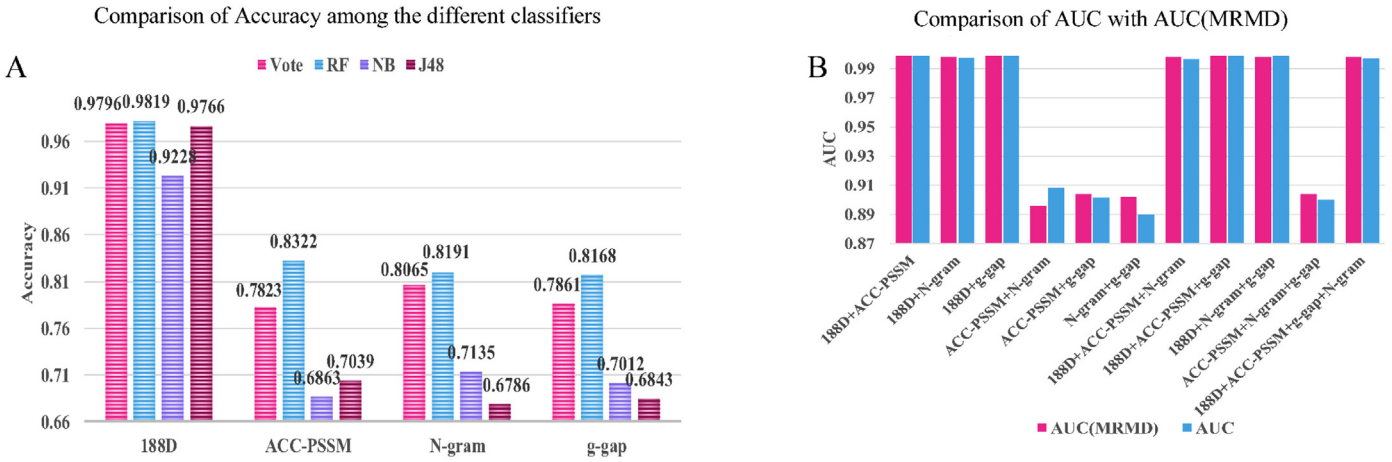

Fig. 3. The result of comparison of evaluation metrics. A. Comparison of Accuracy among the different classifiers; B, Comparison of AUC with AUC (MRMD).

Table 2
The classification results of different hybrid features.

Method	Classifier	AUC	ACC	F-Measure	SN	SP
188D + ACC-PSSM	RF	0.9990	0.9831	0.9833	0.9907	0.9753
188D + N-gram	RF	0.9977	0.9808	0.9810	0.9890	0.9723
188D + g-gap	RF	0.9990	0.9816	0.9813	0.9907	0.9723
ACC-PSSM + N-gram	RF	0.9083	0.8326	0.8320	0.7650	0.9000
ACC-PSSM + g-gap	RF	0.9013	0.8333	0.8327	0.7687	0.8977
N-gram + g-gap	RF	0.8897	0.8168	0.8167	0.7880	0.8457
188D + ACC-PSSM + N-gram	RF	0.9967	0.9700	0.9700	0.9603	0.9800
188D + ACC-PSSM + g-gap	RF	0.9990	0.9831	0.9833	0.9900	0.9763
188D + N-gram + g-gap	RF	0.9987	0.9820	0.9820	0.9857	0.9787
ACC-PSSM + N-gram + g-gap	RF	0.9000	0.8314	0.8310	0.7630	0.9000
188D + ACC-PSSM + g-gap + N-gram	RF	0.9970	0.9689	0.9690	0.9550	0.9820

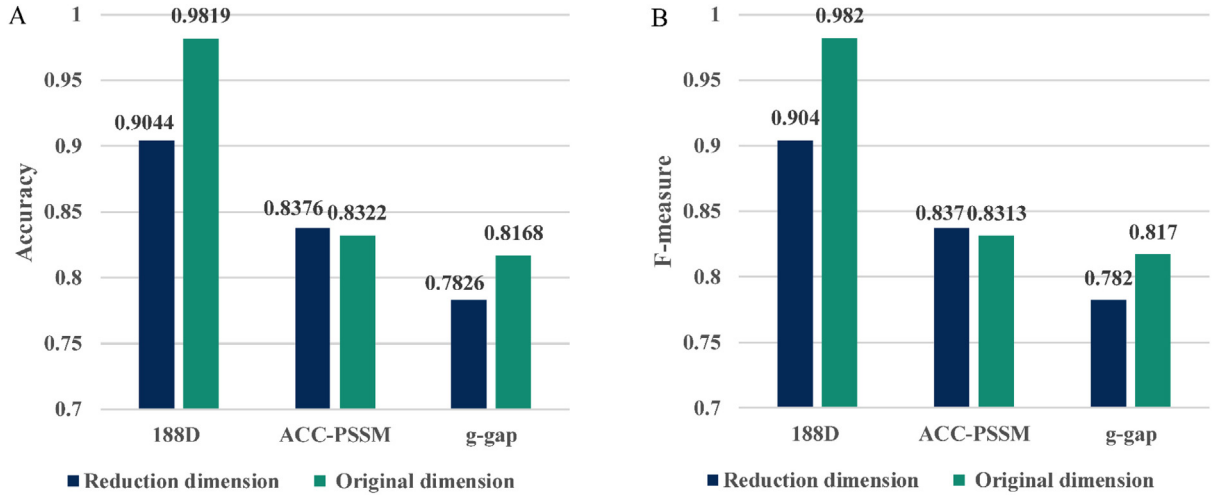

Fig. 4. The ACC and F-measure comparison results. A, Comparison of ACC before and after optimal feature reduction; B, Comparison of F-measure before and after optimal feature reduction.

feature set with fewer but more expressive features. Finally, through principal component analysis, we obtained the 10-dimensional optimal feature set, which was classified and visualized. Similarly, the optimal feature set method is processed for the ACC-PSSM and g-gap features. N-gram features are not suitable for optimal feature set methods because the original dimensions are too large. The results are shown in Fig. 4 and Fig. 5.

The obtained optimal feature of dimension reduction is classified by RF and evaluated by eight-fold cross-validation. The accuracy and F-measure of 188D and g-gap features are decreased, but the accuracy and F-measure of ACC-PSSM are increased. The accuracy and F-measure of the 188D feature have decreased by 7.75% and 7.8%, respectively, and the accuracy and F-measure of the g-gap feature have decreased by 3.42% and 3.5%, respectively (Fig. 4). The prediction performance of the ACC-PSSM feature has improved, and the Accuracy and F-measure have improved by 0.54% and 0.47%, respectively (Fig. 4). Then, we obtained datasets of three optimal feature vectors by the optimal feature set method using t-SNE for dimensionality reduction and visualization. For the 188D feature, the optimal feature is better than the original feature, and the positive samples and negative samples are completely separated (Fig. 5A and B). The comparison results of the other two features (ACC-PSSM and g-gap) are shown in Fig. 5C, D, E and F. The positive samples and negative samples in the visualization are not completely separated. In summary, the 188D feature is the best feature for identifying antioxidant proteins.

3.4. Comparison with other methods

For fair comparison, when comparing with other methods, our proposed new method used the same data as other methods. The new method used hybrid features and random forests to predict antioxidant proteins. The hybrid feature selects the best combination of 188D and ACC-PSSM. Our proposed original method has better predictive performance than other methods (Table 3). The ACC obtained by the proposed method is 9.12% and 17.03% higher than AodPred [33] and Naïve Bayes [29], respectively (Table 3). The experimental results show that this novel method has high accuracy, which can accurately separate antioxidant proteins from non-antioxidant proteins.

4. Conclusion

Protein prediction mainly involves two aspects, feature extraction and selection of classification algorithms. Predicting the performance of antioxidant proteins is related to protein expression and classifiers. In the experiments, we have proposed a new method, which is a predictor to distinguish antioxidant from non-antioxidant proteins based on a combination of hybrid features and random forest. To find the best features of antioxidant proteins, four individual feature extraction methods, including 188D, ACC-PSSM, N-gram, and g-gap, were used to extract the features of antioxidant protein sequences. These four features were based on sequence information, physiochemical properties and correlation. Among the single feature extraction methods, the AUC, ACC, SN, SP, and F-measure obtained using 188D features and RF as the classifier are the highest. Then, the four features are randomly combined using the hybrid feature representation method. The experimental results show that all the indicators obtained from the combination of 188D features and other three features are above 0.9550. Next, we used three feature selection methods, MRMD, t-SNE, and the

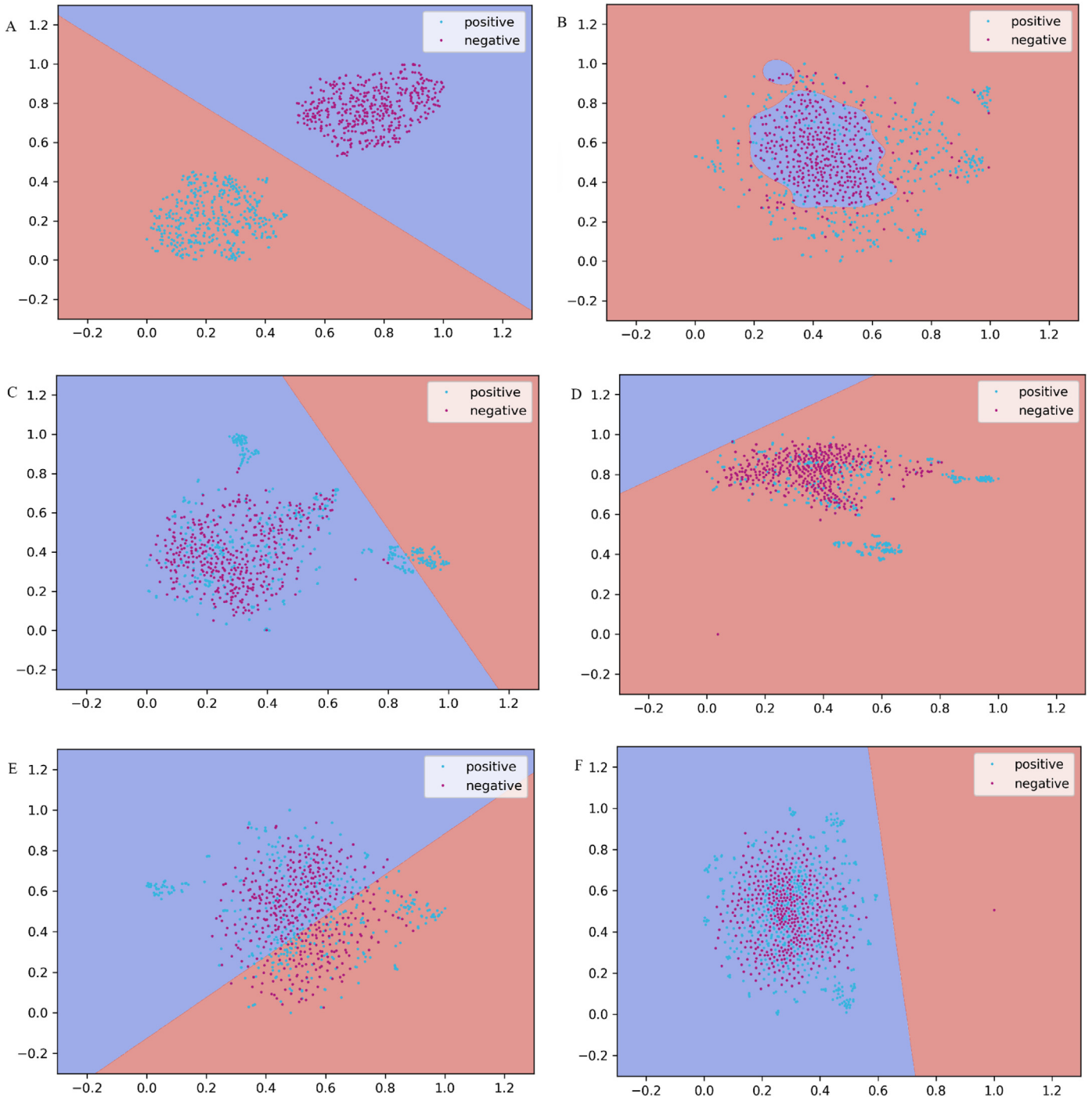

Fig. 5. The results of different optimal feature sets are compared through t-SNE dimensionality reduction and visualization. (A) 188D optimal feature; (B) 188D original feature; (C) ACC-PSSM optimal feature; (D) ACC-PSSM original feature; (E) g-gap optimal feature; (F) g-gap original feature.

Table 3
Comparison of accuracy with other methods.

Method	ACC	SN	SP
Hybird feature (This paper)	**0.8391**	0.665	**0.9630**
AodPred	0.7479	0.7509	0.7448
Naïve Bayes	0.6688	0.7204	0.6605

Bold letters represent the maximum value in each comparison.

optimal feature set selection method, to improve the performance of predicting antioxidant proteins. According to the experimental results, the hybrid feature representation methods are better than the single feature expression method in predicting the performance of antioxidant proteins. The new method compares performance with other methods on the same positive samples and negative samples. The results show that the new method results in improved accuracy, which is 9.12% and 17.03% higher than that of AodPred and Naïve Bayes, respectively. In future work, we will establish an online service website as well as extend the work to other areas. Deep learning [98–100] and other artificial intelligence methods [101–104] are also worthy to test if the data become more enough.

Author's contribution

C.Y.A. and L.Y. conceived and designed the experiment. C.Y.A., W.Y.Z, and L.G. performed the experiment. C.Y.A., W.Y.Z, L.G., B.Z.D and L.Y. analyzed the results. C.Y.A. and W.Y.Z. wrote the manuscript. L.G., B.Z.D and L.Y. revised the manuscript. B.Z.D and L.Y. approved the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Declaration of Competing Interest

All authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910403), National Natural Science Foundation of China (No. 61672406) and the Fundamental Research Funds for the Central Universities [No. JB180307].

References

[1] M.T. Satué-Gracia, et al., Lactoferrin in infant formulas: effect on oxidation, J. Agric. Food Chem. 48 (10) (2000) 4984–4990.
[2] Y.-W. Liu, et al., Patatin, the tuber storage protein of potato (*Solanum tuberosum* L.), exhibits antioxidant activity in vitro, J. Agric. Food Chem. 51 (15) (2003) 4389–4393.
[3] B. Li, et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res. 45 (W1) (2017) W162–W170.
[4] J. Tang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2) (2019) 621–636.
[5] A. Podsędek, Natural antioxidants and antioxidant capacity of Brassica vegetables: a review, LWT Food Sci. Technol. 40 (1) (2007) 1–11.
[6] R.J. Elias, S.S. Kellerby, E.A. Decker, Antioxidant activity of proteins and peptides, Crit. Rev. Food Sci. Nutr. 48 (5) (2008) 430–441.
[7] A.M. Shah, K.M. Channon, Free radicals and redox signalling in cardiovascular disease, Heart 90 (5) (2004) 486–487.
[8] L.A. Pham-Huy, H. He, C. Pham-Huy, Free radicals, antioxidants in disease and health, Int. J. Biomed. Sci. 4 (2) (2008) 89–96.
[9] E. Fernández-Blanco, et al., Random Forest classification based on star graph topological indices for antioxidant proteins, J. Theor. Biol. 317 (2013) 331–337.
[10] B. Ames, Dietary carcinogens and anticarcinogens, Oxygen Radicals Degenerative Dis. 221 (4617) (1983) 1256–1264.
[11] B.N. Ames, M.K. Shigenaga, T.M. Hagen, Oxidants, antioxidants, and the degenerative diseases of aging, 90 (1993), pp. 7915–7922 17.
[12] M. Li, et al., Efficient mini-batch training for stochastic optimization, in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, Association for Computing Machinery, New York, New York, USA, 2014, pp. 661–670.
[13] B. Halliwell, Free radicals, antioxidants, and human disease: curiosity, cause, or consequence? Lancet 344 (8924) (1994) 721–724.
[14] M. Valko, et al., Free radicals, metals and antioxidants in oxidative stress-induced cancer, Chem. Biol. Interact. 160 (1) (2006) 1–40.
[15] L. Jiang, et al., FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association, BMC Genomics 19 (911) (2019) 11–25.
[16] L. Jiang, et al., MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association, Front. Genet. 9 (2018).
[17] L. Yu, F. Xu, L. Gao, Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression, Front Bioeng Biotechnol 8 (2020) 8.
[18] V. Staudacher, et al., Redox-sensitive GFP fusions for monitoring the catalytic mechanism and inactivation of peroxiredoxins in living cells, Redox Biol. 14 (2018) 549–556.
[19] M. Alfonso-Prieto, et al., The molecular mechanism of the catalase reaction, J. Am. Chem. Soc. 131 (33) (2009) 11751–11761.
[20] W. Huang, et al., Purification and characterization of an antioxidant protein from *Ginkgo biloba* seeds, Food Res. Int. 43 (1) (2010) 86–94.
[21] W. Chen, et al., iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68.
[22] S. Lee, B.C. Lee, D. Kim, Prediction of protein secondary structure content using amino acid composition and evolutionary information, Proteins 62 (4) (2006) 1107–1114.
[23] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2004) 10–19.
[24] Z. Lv, et al., A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features, Front. Bioeng. Biotechnol. 7 (2019) 215.
[25] G. Liu, et al., Circulating vitamin E levels and Alzheimer's disease: a Mendelian randomization study, Neurobiol Aging 72 (2018) 189 e1–189 e9.
[26] H. Wang, et al., Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion, Neurocomputing 383 (2020) 257–269.
[27] Y. Shen, J. Tang, F. Guo, Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC, J. Theor. Biol. 462 (2019) 230–239.
[28] Y. Shen, et al., Critical evaluation of web-based prediction tools for human protein subcellular localization, Brief. Bioinform. (2019), https://doi.org/10.1093/bib/bbz106.
[29] P.M. Feng, H. Lin, W. Chen, Identification of antioxidants from sequence information using naive Bayes, Comput Math Methods Med 2013 (2013) 567529.
[30] B. Liu, et al., iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach, Bioinformatics 34 (22) (2018) 3835–3842.
[31] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2) (2012) 629–641.
[32] A.H. Butt, N. Rasool, Y.D. Khan, Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC, J. Theor. Biol. 473 (2019) 1–8.
[33] P. Feng, W. Chen, H. Lin, Identifying antioxidant proteins by using optimal dipeptide compositions, Interdiscip. Sci. 8 (2) (2016) 186–191.
[34] L. Shao, et al., Identification of antioxidant proteins with deep learning from sequence information, Front. Pharmacol. 9 (2018) 1036.
[35] L. Xu, et al., SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins, Int. J. Mol. Sci. (2018) 19(6).
[36] L. Wei, et al., Enhanced protein fold prediction method through a novel feature extraction technique, IEEE Trans. NanoBioscience 14 (6) (2015) 649–659.
[37] C. Meng, et al., AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine, Front Bioeng Biotechnol 7 (2019) 224.
[38] L. Zhang, et al., Incorporating g-gap dipeptide composition and position specific scoring matrix for identifying antioxidant proteins, 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), 2015.
[39] L. Zhang, et al., Sequence based prediction of antioxidant proteins using a classifier selection strategy, PLoS One 11 (9) (2016) e0163274.
[40] Q. Zou, et al., Sequence clustering in bioinformatics: an empirical study, Brief. Bioinform. 21 (1) (2020) 1–10.
[41] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (13) (2006) 1658–1659.
[42] J. Zhang, B. Liu, A review on the recent developments of sequence-based protein feature extraction methods, Curr. Bioinforma. 14 (3) (2019) 190–199.
[43] P.M. Feng, L. Hao, C. Wei, Identification of antioxidants from sequence information using Naïve Bayes, Comput Math Methods Med 2013 (2013) 1–5.
[44] 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 (10) (2010) 1330–1334.
[45] J. Dongardive, S. Abraham, Protein sequence classification based on n-gram and k-nearest neighbor algorithm, in Computational Intelligence in Data Mining—Volume 2, Springer, 2016, pp. 163–171.
[46] L. Xu, et al., SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins, Int. J. Mol. Sci. 19 (6) (2018) 1773.
[47] Y. Zhao, F. Wang, L. Juan, MicroRNA promoter identification in Arabidopsis using multiple histone markers, Biomed. Res. Int. 2015 (2015) 861402.
[48] Q.H. Jiang, et al., Predicting human microRNA-disease associations based on support vector machine, Int. J. Data Mining Bioinform. 8 (3) (2013) 282–293.
[49] L. Cheng, et al., LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse, Nucleic Acids Res. 47 (D1) (2019) D140–D144.
[50] C.Z. Cai, et al., SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (13) (2003) 3692–3697.
[51] Q. Zou, et al., BinMemPredict: a web server and software for predicting membrane protein types, Curr. Proteomics 10 (1) (2013) 2–9.
[52] T. Liu, et al., Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, Amino Acids 42 (6) (2012) 2243–2249.
[53] B. Liu, BioSeq-analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, Brief. Bioinform. 20 (4) (2017) 1280–1294.
[54] J. Wang, et al., POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, Bioinformatics 33 (17) (2017) 2756–2758.
[55] G. Wang, et al., MeDReaders: a database for transcription factors that bind to methylated DNA, Nucleic Acids Res. 46 (D1) (2018) D146–D151.
[56] B. Liu, H. Wu, K.-C. Chou, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (4) (2017) 67.
[57] Q. Dong, S. Zhou, J. Guan, A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation, Bioinformatics 25 (20) (2009) 2655–2662.
[58] Y. Guo, et al., Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic Acids Res. 36 (9) (2008) 3025–3030.
[59] B. Liu, et al., Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning, IEEE Trans. Nanobiosci. 15 (4) (2016) 328–334.
[60] P.F. Brown, et al., Class-based n-gram models of natural language, Computational Linguistics 18 (4) (1992) 467–479.

[61] C. Leslie, E. Eskin, W.S. Noble, The spectrum kernel: a string kernel for SVM protein classification, Pac. Symp. Biocomput. (2002) 564–575.
[62] H. Ding, et al., Prediction of Golgi-resident protein types by using feature selection technique, Chemom. Intell. Lab. Syst. 124 (2013) 9–13.
[63] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[64] X. Chen, et al., Novel human miRNA-disease association inference based on random Forest, Molecular Therapy-Nucleic Acids 13 (2018) 568–579.
[65] C.-C. Wang, et al., RFSMMA: a new computational model to identify and prioritize potential small molecule-MiRNA associations, J. Chem. Inf. Model. 59 (4) (2019) 1668–1679.
[66] G. Nimrod, et al., Identification of DNA-binding proteins using structural, electrostatic and evolutionary features, J. Mol. Biol. 387 (4) (2009) 1040–1053.
[67] X. Sun, et al., RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net, Chemom. Intell. Lab. Syst. 197 (2020) 103919.
[68] D. Yu, et al., Disulfide connectivity prediction based on modelled protein 3D structural information and random Forest regression, IEEE/ACM Trans. Computat. Biol. Bioinform. 12 (3) (2015) 611–621.
[69] Y. Ding, J. Tang, F. Guo, Identification of protein-ligand binding sites by sequence information and ensemble classifier, J. Chem. Inf. Model. 57 (12) (2017) 3149–3161.
[70] X. Zhao, et al., ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles, BMC Bioinformatics 21 (1) (2020) 43.
[71] A. Liaw, M. Wiener, Classification and regression by RandomForest, Forest 23 (2001).
[72] Q. Zou, et al., A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.
[73] K. Pearson, Determination of the coefficient of correlation, Science 30 (757) (1909) 23–25.
[74] C.R. Maurer, Q. Rensheng, V. Raghavan, A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2) (2003) 265–270.
[75] Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* J. Mach. Learn. Res., 2008. 9(Nov): p. 2579–2605.
[76] D. Zhang, M. Kabuka, Protein Family Classification from Scratch: A CNN based Deep Learning Approach, IEEE/ACM Trans. Comput. Biol. Bioinform. (2020), https://doi.org/10.1109/TCBB.2020.2966633.
[77] H. Zhou, F. Wang, P. Tao, T-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations, J. Chem. Theory Comput. 14 (11) (2018) 5499–5510.
[78] X. Zeng, et al., Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (2020) 1775–1797.
[79] Song, B., et al., *Cell-like P systems with evolutional symport/antiport rules and membrane creation.* Inf. Comput., 2020: p. 104542.
[80] C. Feng, Q. Zou, D. Wang, Using a low correlation high Orthogonality feature set and machine learning methods to identify plant Pentatricopeptide repeat coding gene/protein, Neurocomputing (2020), https://doi.org/10.1016/j.neucom.2020.02.079.
[81] B. Cooil, R.S. Winer, D.L. Rados, Cross-Validation for Prediction, 24 (1987), pp. 271–279 3.
[82] J. Chen, et al., Prediction of Protein Ubiquitination Sites in *Arabidopsis thaliana*, 14 (2019), pp. 614–620 7.
[83] S. Jemimah, M. Sekijima, M.M. Gromiha, ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification, Bioinformatics 36 (6) (2019) 1725–1730.
[84] X.-J. Zhu, et al., Predicting protein structural classes for low-similarity sequences by evaluating different features, Knowl.-Based Syst. 163 (2019) 787–793.
[85] M. Stock, et al., Algebraic shortcuts for leave-one-out cross-validation in supervised network inference, Brief. Bioinform. 21 (1) (2018) 262–271.
[86] L. Xu, et al., k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification, Front. Genet. (2019) 10(33).
[87] L. Xu, et al., An efficient classifier for Alzheimer's disease genes identification, Molecules 23 (12) (2018) 3140.
[88] C. Shen, et al., LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information, IEEE Access 7 (2019) 13486–13496.
[89] Q. Jiang, et al., Prioritization of disease microRNAs through a human phenome-microRNAome network, BMC Syst Biol 4 (Suppl. 1) (2010) S2.
[90] L. Cheng, et al., DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function, Bioinformatics 34 (11) (2018) 1953–1956.
[91] X. Zeng, et al., A consensus community-based particle swarm optimization for dynamic community detection, IEEE Trans. Cybernetics (2019), https://doi.org/10.1109/TCYB.2019.2938895.
[92] Z. Hong, et al., Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism, Bioinformatics 36 (4) (2019) 1037–1043.
[93] X. Zeng, et al., Network-based prediction of drug-target interactions using an arbitrary-order proximity embedded deep Forest, Bioinformatics 36 (9) (2020) 2805–2812.
[94] G. Wang, et al., Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells, BMC Genom 9 (Suppl. 2) (2008) S22.
[95] Y. Zhao, et al., Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network, Biomed. Res. Int. 2017 (2017) 7049406.
[96] G. Wang, et al., Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells, PLoS One 5 (7) (2010) e11794.
[97] N. Williams, S. Zander, G. Armitage, A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification, ACM SIGCOMM Comp. Commun. Rev. 36 (5) (2006) 5–16.
[98] L. Peng, et al., The advances and challenges of deep learning application in biological big data processing, Curr. Bioinforma. 13 (4) (2018) 352–359.
[99] Z.B. Lv, C.Y. Ao, Q. Zou, Protein function prediction: from traditional classifier to deep learning, Proteomics 19 (14) (2019) 2.
[100] L. Yu, et al., Drug and nondrug classification based on deep learning with various feature selection strategies, Curr. Bioinforma. 13 (3) (2018) 253–259.
[101] H. Xu, et al., An evolutionary algorithm based on Minkowski distance for many-objective optimization, IEEE Trans. Cybernetics 49 (11) (2019) 3968–3979.
[102] H. Xu, et al., MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition, IEEE Trans. Cybernetics 49 (2) (2019) 517–526.
[103] T. Song, et al., Spiking neural P systems with Colored spikes, IEEE Trans. Cognit. Develop. Syst. 10 (4) (2018) 1106–1115.
[104] F.G.C. Cabarle, et al., On solutions and representations of spiking neural P systems with rules on synapses, Inf. Sci. 501 (2019) 30–49.