frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 29 October 2020
doi: 10.3389/fcell.2020.591487

Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions

*Yixiao Zhai†, Yu Chen†, Zhixia Teng and Yuming Zhao**

Information and Computer Engineering College, Northeast Forestry University, Harbin, China

Excessive oxidative stress responses can threaten our health, and thus it is essential to produce antioxidant proteins to regulate the body's oxidative responses. The low number of antioxidant proteins makes it difficult to extract their representative features. Our experimental method did not use structural information but instead studied antioxidant proteins from a sequenced perspective while focusing on the impact of data imbalance on sensitivity, thus greatly improving the model's sensitivity for antioxidant protein recognition. We developed a method based on the Composition of k-spaced Amino Acid Pairs (CKSAAP) and the Conjoint Triad (CT) features derived from the amino acid composition and protein-protein interactions. SMOTE and the Max-Relevance-Max-Distance algorithm (MRMD) were utilized to unbalance the training data and select the optimal feature subset, respectively. The test set used 10-fold crossing validation and a random forest algorithm for classification according to the selected feature subset. The sensitivity was 0.792, the specificity was 0.808, and the average accuracy was 0.8.

Keywords: antioxidant protein, unbalanced dataset, random forest, machine learning, sequence feature

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Guang Song,
Johns Hopkins University, United States
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Yuming Zhao
zym@nefu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *04 August 2020*
Accepted: *18 September 2020*
Published: *29 October 2020*

Citation:
Zhai Y, Chen Y, Teng Z and Zhao Y (2020) Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. Front. Cell Dev. Biol. 8:591487. doi: 10.3389/fcell.2020.591487

INTRODUCTION

Reactive oxygen species (ROS) are products of metabolic processes (Birben et al., 2012) and include singlet oxygen, hydrogen peroxide, nitric oxide, superoxide anion radicals, and hydroxyl radicals. Excessive concentrations of ROS produce excessive oxygen radicals, and the antioxidant system in the organism cannot eliminate the ROS quickly enough, which causes oxidative stress (OS) (Schieber and Chandel, 2014). An excessive OS response can affect the destruction of the macromolecular structure, such as the DNA, proteins and other carbohydrates, and even give rise to cell death, which can lead to aging (Liguori et al., 2018) and initiate genetic diseases. At present, people have realized that the OS response has a role in the pathogenesis of many diseases, including cancer, acute and chronic kidney diseases, neurodegenerative diseases, cardiovascular disease, diabetes, and atherosclerosis (Pisoschi and Pop, 2015; Liguori et al., 2018).

In order to prevent excessively high concentrations of ROS from causing cell damage, the antioxidant proteins must be employed to strike a good balance between the oxidation process and the antioxidant process, which is essential. Given that antioxidant proteins have such powerful functions, accurate identification of antioxidant proteins is absolutely critical for revealing the deterioration of tissue function caused by certain diseases and aging, and for developing new types of antioxidant drugs that can treat or mitigate these types of diseases. However, traditional methods for identifying antioxidant proteins have the problems of being time-consuming and costly, such as western blots (Mahmood and Yang, 2012).

With the continuous improvement of genomic data (Xu et al., 2017; Wang et al., 2018; Zhou et al., 2018; Guo and Zou, 2019; Wang J. et al., 2020), sequencing technology and computer technology, data mining and machine learning methods (Quan et al., 2017; Zou et al., 2017) are being exploited to identify antioxidant proteins, and many researchers have already done so. In Feng et al. (2013) proposed an idea using Naive Bayes, based on sequence information, and after 3 years, they changed the method of data processing and proposed a model called AodPred (Feng et al., 2016). It was based on a support vector machine with 3-spaced residue pairs and its accuracy was significantly better than the former model. In 2016, an integration method was proposed by Zhang et al. (2016), which was applied for predicting antioxidant proteins with mixed features, indicating that protein secondary structure information facilitates the discrimination of target proteins. Then, a method called SeqSVM was presented by Xu et al. (2018) employing a 188D feature extraction method. Last year, Meng et al. (2019) also utilized a support vector machine with structural features to establish a model to discriminate target proteins.

Despite the strengths of the existing methods, there are still some shortcomings that have not been fully addressed. (1) Most methods did not consider the impact of data unbalances on classification when training samples. The feature subset after feature selection was more representative of the larger number of type (non-antioxidant proteins), and what we require to find is a feature subset that is more representative of antioxidant proteins. For example, in Meng's experiment, the sensitivity and specificity of the test set results were 0.68 and 0.985, which meant that the sorted features were more conducive to the selection of non-antioxidant proteins. These problems also existed in Xu's research, even if she did use an unbalanced treatment. (2) Features of protein secondary structure information are extracted based on the secondary structure predicted by sequence information using tools such as PSI-PRED (McGuffin et al., 2000). The whole process is complicated and time-consuming. In addition, there are errors in the predicted protein secondary structures, which also affect the accuracy of the features.

To address the above limitations and to enhance the predictive performance of the antioxidant proteins, the protein was described based on its hybrid features without structural information, including the Composition of k-spaced Amino Acid Pairs (CSKAAP) and the Conjoint Triad (CT) features. At the same time, taking into account the unbalanced state of the data volume of antioxidant proteins and non-antioxidant proteins, oversampling, under-sampling, and combined methods were used to process the dataset. The Max-Relevance-Max-Distance algorithm (MRMD) (Zou et al., 2016) could be exploited to single out the best feature subset for reducing the computational complexity and noise. On the contrary, we chose a 10-fold crossing test and random forest as the classifier, which has the characteristics of a fast running speed and less overfitting, rather than the very popular support vector machine. **Figure 1** shows the complete data processing approach.

MATERIALS AND METHODS

Benchmark Dataset

The dataset we used has been previously used by Feng et al. (2016), Xu et al. (2018), and Meng et al. (2019). We first collected proteins with antioxidant activities from the UniProt database (release 2014_02) according to the following steps: (1) only proteins with experimentally proven antioxidant activities were selected; and (2) ambiguous proteins were excluded, such as those containing non-standard letters like "B," "X," and "Z." After this rigorous screening, we obtained 710 protein sequences as the original positive samples for the experiment. The negative samples were 1567 PDB proteins with identification values $<$20%, which were picked by PISCES-culled. To reduce redundancy and to avoid homology bias (Zou et al., 2020), peptides with more than 60% sequence similarity to each other were removed from the benchmark dataset by the CD-HIT program. Finally, the new dataset, including 1805 proteins sequences, was obtained, and 253 were antioxidant proteins and 1552 were non-antioxidant proteins. This can be expressed as follows:

$$Dataset = Dataset_{+} \cup Dataset_{-} \quad (1)$$

Where Dataset_{+} indicates the positive dataset, which contains 253 antioxidant proteins; Dataset_{-} indicates 1552 non-antioxidant proteins as the negative dataset; and the "$\cup$ " represents the symbol of "union" in the set theory, which means the benchmark dataset consisted of Dataset_{+} and $Dataset_{-}$. The proportion of antioxidant to non-antioxidant samples is $\sim$1:6, which shows this is an unbalanced dataset.

As we all know, an unbalanced number of positive and negative samples will affect the accuracy. In order to prevent this from happening and to enhance the precision, 200 antioxidant and 1500 non-antioxidant proteins from the final benchmark dataset were selected as the training dataset and the rest with 53 antioxidant proteins and 52 non-antioxidant proteins was set as an independent testing dataset. The next section will detail how we deal with unbalanced training sets.

Feature Extraction

The secondary structure information of the protein takes a long time to extract and process, and the calculation is complicated. In order to simplify the process, and at the same time, considering the diversity and complexity of the function of the antioxidant protein itself, mixture features were adopted to represent antioxidant proteins, including CKSAAP and CT. The CKSAAP describes the composition of amino acids. The other is a feature that describes protein-protein interaction (PPI) information (Yu et al., 2010, 2020; Liu et al., 2019b; Zhao et al., 2020). Three adjacent amino acids are regarded as a linker to judge the charge properties and hydrophobicity of the target protein. The iFeature was employed to extract features, which is a python toolkit. Assuming that a protein sequence consists of N amino acids, where A_i is the ith amino acid in the sequence, it can be defined as:

$$P = A_1A_2A_3, \cdots, A_N \quad (2)$$

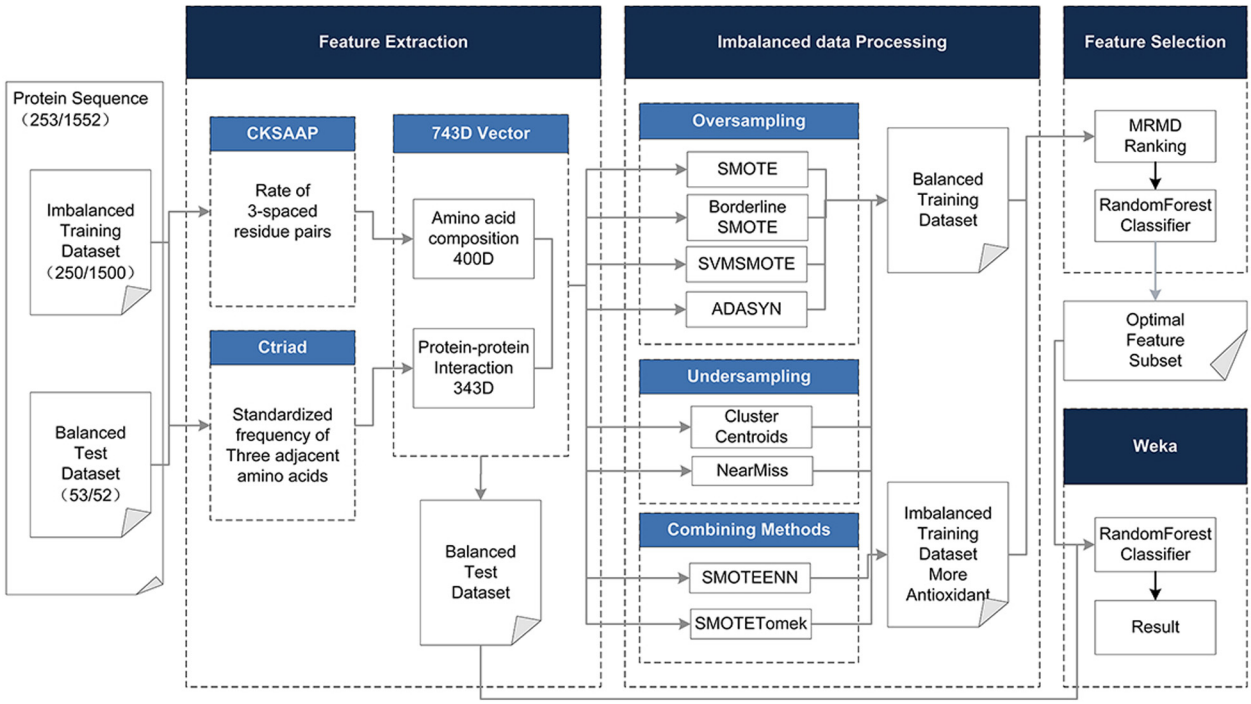

FIGURE 1 | The method flowchart. The original dataset (training and test dataset) is processed in four phases. (1) Using CKSAAP and CT to extract 743D features. (2) In the unbalanced data processing phase, eight methods are adopted to balance the training dataset. (3) In the feature selection phase, the 743D features by MRMD score are ranked and the optimal feature set is selected by Random Forest classifier. (4) Use the selected feature subset to classify the test set to get the final result.

Composition of k-Spaced Amino Acid Pairs

The Composition of k-spaced Amino Acid Pairs (CKSAAP) feature delegates the component of amino acids (Tan et al., 2019; Liu et al., 2020). It calculates on behalf of the frequency of two amino acids separated by k residues (Chen et al., 2007a,b, 2008, 2009). Feng et al. (2016) has confirmed that a 3-spaced residue pairs feature is beneficial for classifying antioxidant proteins, and thus we only chose k=3 in our research, which picked up 400 dimensions. The 20 kinds of amino acids are combined in pairs to get 400 amino acid pairs. We can count the frequency of 400 amino acid pairs in a protein sequence. Then, a 3-spaced feature vector can be defined as:

$$CKSAAP = \left[f_1, f_2, f_3, \cdots, f_{400}\right]^T \tag{3}$$

where the T is the transpose of the *CKSAAP* vector and f_i is the frequency of the ith amino acid pair, which is defined as:

$$f_i = \frac{n_i}{N-4} \tag{4}$$

where n_i is the number of times the ith amino acid pair appears in a protein sequence and N is the length of the sequence. The value of $N-4$ represents the number of 3-spaced amino acid pairs in the whole protein sequence.

TABLE 1 | Classification of amino acids.

No.	*Dipole scale*[a]	*Volume scale*[b]	Class
1	−	−	Ala, Gly, Val
2	−	+	Ile, Leu, Phe, Pro
3	+	+	Tyr, Met, Thr, Ser
4	++	+	His, Asn, Gln, Tpr
5	+++	+	Arg, Lys
6	+′+′+′	+	Asp, Glu
7	+[c]	+	Cys

[a] *Dipole scale (Debye): −, Dipole < 1.0; +, 1.0 < Dipole <2.0; ++, 2.0 < Dipole < 3.0; +++, Dipole > 3.0; +′+′+′, Dipole > 3.0 with opposite orientation.* [b] *Volume scale (Å^3): −, Volume < 50; +, Volume > 50.* [c] *Cys is separated from class 3 because of its ability to form disulfide bonds.*

Conjoint Triad

The Conjoint Triad descriptor (CT) describes the important information of protein-protein interactions (PPI). It is based on the triplet formed between amino acids and adjacent amino acids as the basic unit, considering the connections among them (Shen et al., 2007). First, by measuring the size and side chain volume of each amino acid dipole, and the effect of synonymous mutations, it classifies the 20 amino acids into seven categories. See **Table 1** for the classification results of the 20 amino acids. According to the classification results and the three adjacent amino acids as a unit of this extraction method, we can use the CT algorithm to extract 343 dimensional features. The detailed definitions and descriptions for the structure of the 343 dimensional features are

illustrated in **Figure 2**. Thus, the CT feature vector can be defined as:

$$CT = \left[d_1, d_2, d_3, \cdots, d_{343}\right]^T \quad (5)$$

where the T is the transpose of CT vector and d_i is the normalized frequency of the ith amino acid triad, which is defined as:

$$d_i = \frac{v_i - min\left\{v_1, v_2, \cdots, v_{343}\right\}}{max\left\{v_1, v_2, \cdots, v_{343}\right\}} \quad (6)$$

where v_i is considered to be the frequency of these different trimmers in the antioxidant protein sequence. Finally, the above features follow the order of CKSAPP followed by CT, thus forming a set *FeatureSet* of 743 features, which can be defined as:

$$FeatureSet = \left[f_1, f_2, f_3, \cdots, f_{400}, d_1, d_2, \cdots, d_{343}\right]^T \quad (7)$$

Unbalanced Data Processing

The unbalanced sample size will cause over-fitting of the sample with a large proportion (Wan et al., 2017; Fdez-Glez et al., 2018; Chao et al., 2019; Cheng et al., 2019; Liu, 2019), that is to say, the prediction is biased toward a classification with a larger number of samples, which will reduce the applicability of the model. The processing method at the data level is sampling. Under-sampling, over-sampling, and combined methods are three common and widely used approaches.

In this work, eight different methods from an unbalanced-learning library (Lemaître et al., 2017) were adopted to deal with the unbalanced data. These eight methods include SMOTE, ADASYN, BorderlineSMOTE, SVMSMOTE, ClusterCentroids, NearMiss, SMOTEENN, and SMOTETomek, which cover the three standard methods described above.

The data sets processed by the above methods were subjected to the same subsequent experimental operations, so as to compare the results obtained by different processing methods, and to select a method that is more suitable for processing antioxidant proteins.

Feature Selection

If the extracted features are directly input into the subsequent classifier without any processing, it is difficult to obtain the ideal results (Wang et al., 2010; Tang et al., 2016, 2018; Basith et al., 2019; Liu and Li, 2019; Manavalan et al., 2019a). Further screening of the features, which can better reflect the characteristics of antioxidant proteins, is necessary. In this study, the Max-Relevance-Max-Distance algorithm (MRMD) was used for noise reduction. It mainly completes two steps, calculating the contribution of each feature to the classification first, and then selecting the best feature subset.

In order to rank all features of the sample, we calculated the MRMD score of each feature, which consists of a relevant value and a distance value. The relevant value is expressed as the relationship between the features and sample categories, calculated using the Pearson correlation coefficient as follows:

$$PCC\left(\vec{X}, \vec{Y}\right) = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k - \bar{x}\right)\left(y_k - \bar{y}\right)}{\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k - \bar{x}\right)^2}\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(y_k - \bar{y}\right)^2}} \quad (8)$$

x_k and y_k are the kth element of $\vec{X}$ and $\vec{Y}$, which are two vectors. $\bar{x}$ and $\bar{y}$ are, respectively, the mathematical expectations of $\vec{X}$ and $\vec{Y}$. And the value of MR (Max-Relevance) for every feature is defined as MR_i. The $\overrightarrow{F_i}$ is a vector that is represented in the ith feature and $\vec{C}$ is a target class vector of each instance.

$$MR_i = \left|PCC\left(\overrightarrow{F_i}, \vec{C}\right)\right| (1 \le i \le M) \quad (9)$$

The distance value measures the independence of every feature. The higher the distance, the greater the independence. The MRMD provides three methods for calculating distance. In our research, the choice is Euclidean distance. We utilized the Euclidean distance to calculate the distance between each feature $\overrightarrow{F_i}$ and the other features, which is defined as follows:

$$ED\left(\overrightarrow{F_i}, \overrightarrow{F_k}\right) = \sqrt{\sum_{k=1}^{N}\left(x_i - x_k\right)^2} \left(1 \le k \le M, k \ne i\right) \quad (10)$$

Then, based on this formula, we can obtain the Euclidean distance value of each feature MD_i, which is the final value of Max-Distance. The larger the MD_i value, the lower the redundancy.

$$MD_i = \frac{1}{M-1}\sum ED\left(\overrightarrow{F_i}, \overrightarrow{F_k}\right) (1 \le i \le M) \quad (11)$$

According to MR_i and MD_i, the MRMD score is defined as:

$$MRMD = MR_i + MD_i \quad (12)$$

The features in the set are sorted from high to low according to the score of *MRMD*. Each time a feature with the highest *MRMD* score is added, the classifier of the random forest is input for sorting, and finally, the feature subset with the highest accuracy and the least feature number is selected.

Random Forest

Random forest (Liaw and Wiener, 2002) is an integrated algorithm that integrates multiple trees through the idea of integrated learning. It has been widely used in bioinformatics (Liu et al., 2019a; Manavalan et al., 2019b,c; Wang et al., 2019; Lv H. et al., 2020; Lv Z. B. et al., 2020; Wang M. et al., 2020). It is composed of N decision trees. After the sample is input into the random forest, each decision tree will get a classification result, then N trees will obtain N classification results. The voting results of all classification results are counted, and the category with the most votes is the final output.

In our study, we adopted the random forest as the classifier because it has several advantages suitable for our data. The feature dimension extracted by the combined method of CKSAAP and CT is very high. Even after dimensionality reduction, it still belongs to high-dimensional data. Random forest can handle

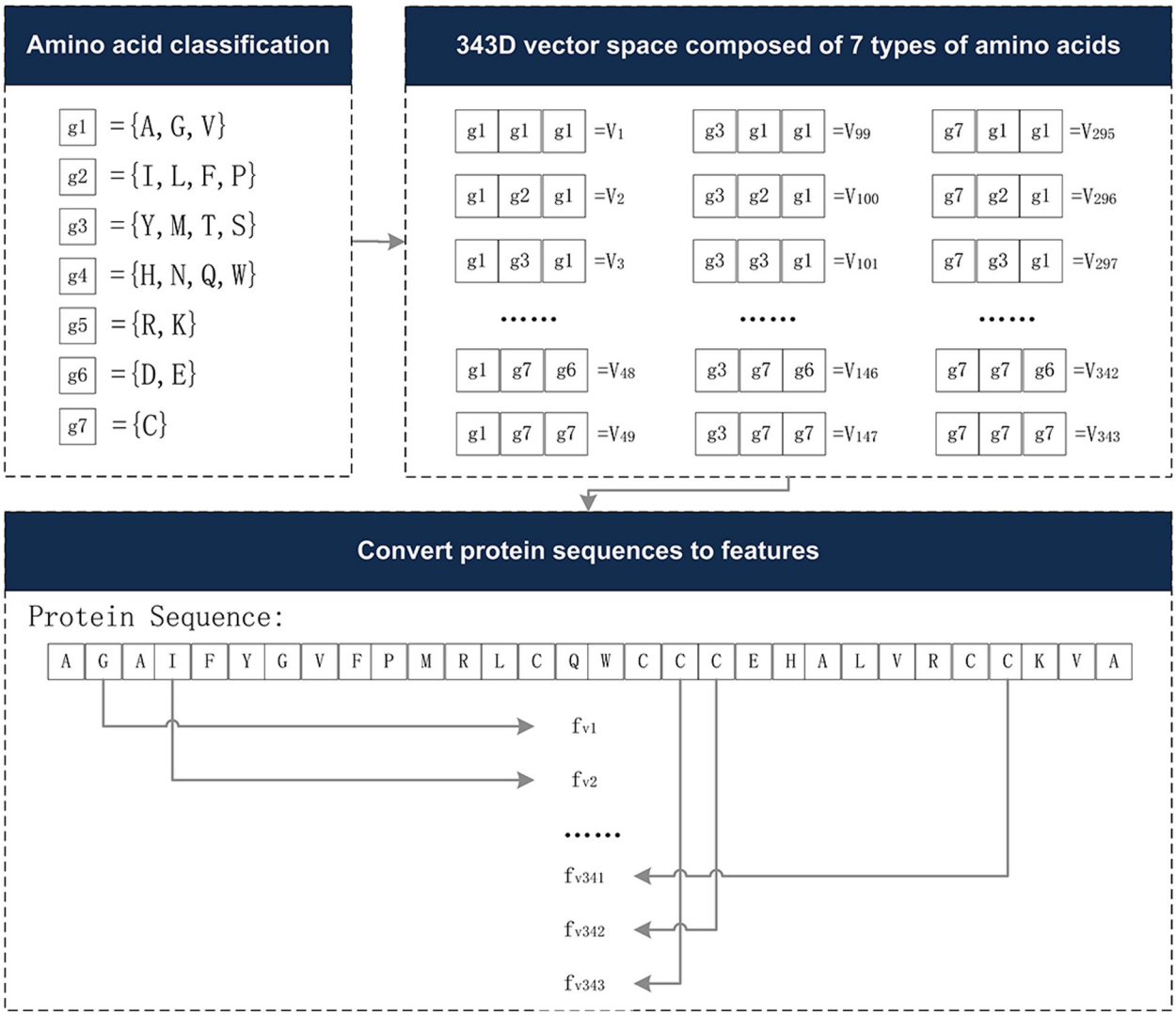

FIGURE 2 | The 343-dimensional feature composition diagram. (1) Classify 20 amino acids into seven categories and obtain the g1∼g7, as shown in the amino acid classification part. (2) In the 343D vector space composed of seven types of amino acids, the seven amino acids were arranged and combined, resulting in 343 trimmers. (3) Examples of sequence conversion into features. The figure was adapted from the Supplementary Figure in Shen et al. (2007).

high-dimensional data, and the accuracy rate is not affected. The training set is unbalanced, and the amount of data becomes larger after the oversampling method is used. Random forest processing is adopted, and the running speed is fast. It is particularly useful in estimating the inferred mapping, so that there is no need to debug many parameters like SVM (Huo et al., 2020).

Measurements

In statistical prediction, there are three commonly used evaluation methods for checking the accuracy of the model (Wang et al., 2008; Basith et al., 2018, 2020; Liu et al., 2019c; Yu et al., 2019; Zhu et al., 2019; Hasan et al., 2020), including the independent dataset sampling test, the k-fold cross validation and the jack-knife test. The jack-knife test is a resampling technique that is suitable for estimating the deviation over the entire sample (Li et al., 2019; Yang et al., 2019). This method has also been used in previous studies, i.e., Feng et al. (2016) and Meng et al. (2019). However, in our study, the training dataset was balanced by oversampling and under-sampling. The training set and test set were mutually exclusive. In order to reduce the complexity of the calculation, 10-fold cross validation is employed. For binary classification problems, the commonly used evaluation indicators are sensitivity (Sn), specificity (Sp), accuracy (Acc), F-score (F), Matthew's Correlation Coefficient (MCC), and the Area Under the Curve (AUC).

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

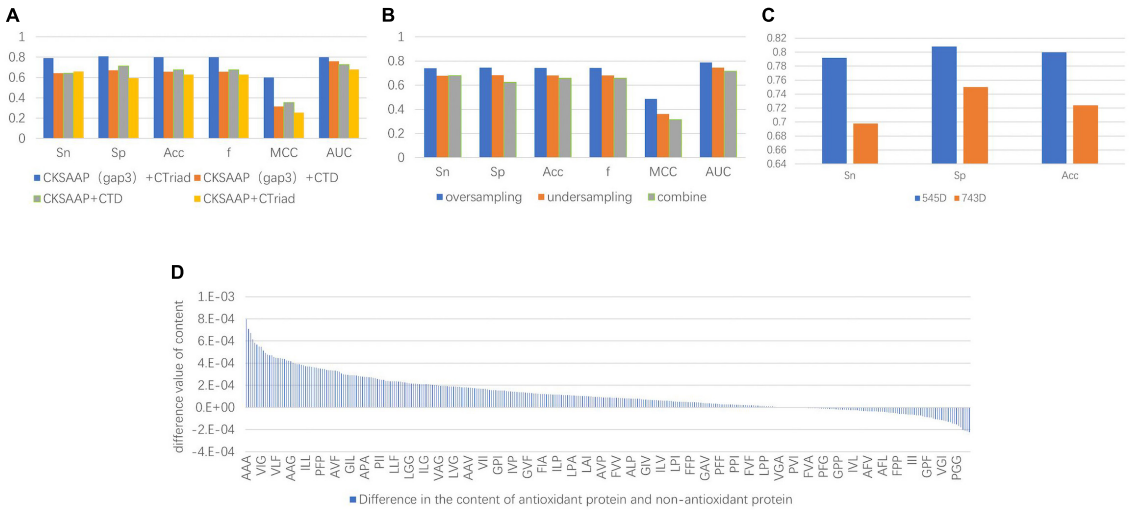

FIGURE 3 | (A) The comparison chart of the results. It was obtained by different feature extraction methods and it shows that the results of the evaluation indicators using the CKSAAP+CT method were higher than the other methods. **(B)** The average value of the final results obtained by using three types of unbalanced processing methods. The result obtained by oversampling is much higher than the other methods. It can be seen that repeated sampling of a small number of sample data to synthesize new data is more conducive to extracting features that make it easy to distinguish antioxidant proteins. **(C)** Comparison of classification effects before and after dimensionality reduction with MRMD. Sn, Sp, and Acc are greatly improved, and the Sn and Sp results are very balanced. **(D)** Compared with sequence characteristics of the antioxidant proteins and non-antioxidant proteins, the triplet of the first type amino acid and the second type amino acid combination appears more frequently.

$$Acc = \frac{TN + TP}{TP + FN + TN + FP} \tag{15}$$

$$F = \frac{2 \times TP}{2TP + FN + FP} \tag{16}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)\,(TP + FP)\,(TN + FP)\,(TN + FN)}} \tag{17}$$

where TP, FP, FN, and TN indicate true positive, false positive, false negative, and true negative, respectively. F is a weighted harmonic average of precision and recall, which can avoid the contradiction between both. MCC is suitable for measuring imbalanced data sets, which is an index used in machine learning to measure the classification performance of two categories. In addition, the AUC is an evaluation index that measures the pros and cons of the binary classification model, which can make a reasonable evaluation of the classifier when the samples are unbalanced (Zhao et al., 2015, 2017; Manavalan et al., 2018a,b; Yu and Gao, 2019; Tang et al., 2020). The larger the AUC value, the better the performance of the model. The value of AUC is the area enclosed by the receiver operating characteristic curve (ROC curve) and the x-axis and y-axis. The vertical axis and the horizontal axis of the ROC curve are Sn and (1-Sp).

RESULTS

Comparison of the Different Feature Extraction Methods

According to existing research, it has been confirmed that a series of feature extraction methods are effective for classifying antioxidant proteins, such as g-gap dipeptides feature, CTD, SSI, RSA, PSSM, etc. Therefore, in the planning stage of the

TABLE 2 | The accuracy rate of eight data imbalance processing methods in different classifiers.

	RF	LIbD3C	LibSVM
Smote	**0.8**	**0.571**	**0.533**
ADASYN	0.705	0.686	0.533
BorderlineSMOTE	0.733	0.638	0.533
SVMSMOTE	0.733	0.705	0.533
ClusterCentroids RandomState = 0	0.667	0.648	0.6
NearMiss version = 1	0.733	0.686	0.6
NearMiss version = 2	0.638	0.648	0.6
NearMiss version = 3	0.686	0.638	0.6
SMOTEENN	0.59	0.571	0.562
SMOTETomek	0.724	0.648	0.543

TABLE 3 | Compared the best results in our research with the results of AodPred.

	Sn	Sp	Acc
AodPred	0.751	0.745	0.748
Smote+RF	**0.792**	**0.808**	**0.800**

TABLE 4 | The prediction result of the model established by different data imbalance processing methods.

	Sn	Sp	Acc	F	MCC	AUC
SMOTE	**0.792**	**0.808**	**0.8**	**0.8**	**0.6**	**0.8**
ADASYN	0.698	0.712	0.705	0.705	0.41	0.766
BorderlineSMOTE	0.736	0.731	0.733	0.733	0.467	0.807
SVMSMOTE	0.736	0.731	0.733	0.733	0.467	0.78
ClusterCentroids RandomState = 0	0.604	0.731	0.667	0.665	0.337	0.743
NearMiss version = 1	0.755	0.712	0.733	0.733	0.467	0.775
NearMiss version = 2	0.66	0.615	0.638	0.638	0.276	0.733
NearMiss version = 3	0.698	0.673	0.686	0.686	0.371	0.734
SMOTEENN	0.604	0.557	0.59	0.59	0.181	0.628
SMOTETomek	0.755	0.692	0.724	0.724	0.448	0.805

experiment, we chose CKSAAP and CTD, and CT combined separately without structure information, and looked for the most suitable feature combinations for the target protein. Among them, CKSAAP was divided into only containing 3-spaced residue pairs and containing *g*-spaced residue pairs (g=1, 2, 3, 4, 5).

In addition, we adopted the principle of a single variable, controlling other factors unchanged, only changing the method of feature extraction, and observed its impact on the experimental results. After the feature extraction was completed, SMOTE and MRMD were used to perform unbalanced processing and to select the optimal feature subset. The final result was obtained by using a random forest classifier and the 10-fold cross-validation method.

The experimental results indicated that the groups only containing 3-spaced residue pairs were superior than the others for classification, which also confirmed the conclusion that the 3-gap dipeptides feature in Feng et al. (2016) was good for classification. On the other hand, despite previous research showing that CTD could be used to obtain good classification results, such as the combined features of Zhang et al. (2016) and Xu et al. (2018) with 188D, in fact, the experimental results showed the classification accuracy of the CT groups was higher than that of the CTD groups. Therefore, only containing 3-spaced residue pairs and CT were selected as the methods of feature extraction. The comparison of the experimental results is shown in **Figure 3A**.

Comparison of the Different Classifier and AodPred

In this experiment, three alternative classifiers were selected, namely LibSVM (Chang and Lin, 2011; Jiang et al., 2013), LibD3C (Lin et al., 2014), and Random Forest. LIBSVM is an SVM pattern recognition and regression software package, which was developed and designed by Prof. Lin Zhiren of Taiwan University. It has the characteristics of a simple to use method, fast operation speed and strong practicability. When using LibSVM, we input the training data into the gird.py file, and entered the values of the calculated parameters c and g into the LibSVM classifier embedded in WEKA (Hall et al., 2009), and then classified the test set. LibDC developed by Lin et al. (2014) is an integrated classifier, which combines multiple basic classification algorithms. Both LibD3C and Random Forest used the embedded WEKA version and we used their default methods to classify the test set.

The classification results showed that the two classification methods of LibSVM and LibD3C had the phenomenon of over-fitting, and the generalization ability of the test set was weak, while when using random forest, the generalization performance of the classification was stronger and more stable. Compared with the existing research AodPred, the method of random forest was higher than AodPred for sensitivity, specificity and accuracy. The comparisons of the experimental results are shown in **Tables 2**, **3**. **Table 2** shows the accuracies of 8 data unbalanced processing methods with different classifiers. **Table 3** compares the best results in our research with the results of the known model AodPred. Our results were obtained after processing using the Smote method, dimensionality reduction using MRMD, and selected features using random forest classifiers applied to the test set.

Comparison of the Different Unbalanced Data Processing Methods

We employed over-sampling, under-sampling and combined methods to deal with the unbalanced training data set. The methods used for oversampling were SMOTE, ADASYN, BorderlineSMOTE, and SVMSMOTE. The parameter settings of each method were the default parameters in the unbalanced library of python. The processed training set samples reached equilibrium, with 1500 positive examples and 1500 negative examples, respectively. ClusterCentroids and NearMiss were the methods of under-sampling. The parameter setting of ClusterCentroids was the default. The version parameters of the NearMiss method take 1, 2, and 3 for unbalanced data processing. Therefore, there were four actual undersampling methods. The processed training data contained 200 positive examples and 200 negative examples. SMOTEENN and SMOTETomek adopted SMOTE to combine with ENN and Tomek, respectively, which were combined methods. In our study, the parameter settings of both were also the default. After SMOTEENN, the processed dataset was also unbalanced, which including 1498 antioxidant proteins and 29 non-antioxidant proteins. Although the processed data was still in an unbalanced state, most of them were antioxidant proteins, which helped us screen out the features with obvious signals. Unlike SMOTEENN, the

data processed by SMOTETomek was balanced, including 1500 positive examples and 1500 negative examples.

After the unbalanced training data, the optimal feature subset was selected by MRMD, and the test set was classified according to the different feature subsets. The experimental results showed that the model obtained by the data processed by the oversampling method had a higher sensitivity (Sn), specificity (Sp), accuracy (Acc), f score (F), Matthew's Correlation Coefficient (MCC), and the Area Under the Curve (AUC) than the other two methods. The reason is that there are fewer antioxidant proteins, and repeated sampling of samples to strengthen their signal characteristics is more conducive to screening out antioxidant proteins. The comparisons of experimental results are shown in **Table 4** and **Figure 3B**. **Table 4** is the prediction results of the model established by different data unbalanced processing methods in the test set. **Figure 3B** shows the average of the prediction results of the models created by the three basic data unbalanced processing methods in the test set.

Feature Contribution and Importance Analysis

The dimension of the original feature was 743D. After the feature selection of MRMD, the selected feature subset contained 545 features. Compared with the original features, the accuracy of the test set classification was improved by 0.076. The experimental results after dimensionality reduction were as follows: the sensitivity was 0.792, the specificity was 0.808, and the average accuracy was 0.8. Compared with the original method, the sensitivity was greatly improved. The comparison chart is shown in **Figure 3C**.

Not only that, by comparing the characteristic MRMD scores, we recognized that CT scores were generally higher than CKSAAP, and the characteristic scores composed of triplets composed of the first and second amino acids in CT were the highest. This means that there were differences in these characteristics between the positive and negative examples. Therefore, we counted the differences in the content of the triplet composed of the first type (A, G, V) and second type (I, L, F, P) of amino acids. There were a total of 343 triplets composed of these amino acids. Among the 260 features, the average content in antioxidant proteins was higher than that of non-antioxidant proteins. The content difference chart is shown in **Figure 3D**. A, V, I, L, F, and P were hydrophobic amino acids, and the tripeptide group composed of them was also hydrophobic, and thus we can infer that the hydrophobicity of proteins can be used to classify antioxidant proteins.

DISCUSSION

In this paper, we proposed a method with CKSAAP and CT features to identify antioxidant proteins. SMOTE was adopted to deal with unbalanced data, and we selected the optional feature set with MRMD. Using the 10-fold cross-validation and random forest classifier on the test set, we obtained an average accuracy of 0.8. The sensitivity and specificity were 0.792 and 0.808, respectively. We revealed that due to the small number of antioxidant proteins, when dealing with an unbalanced problem, oversampling to strengthen the antioxidant proteins makes it easier to discover the signal characteristics that represent the proteins. Therefore, oversampling is more suitable than under-sampling and combination methods. From the experimental results, the SMOTE method works the best. Additionally, after analyzing the characteristics, we found that the sequence of the antioxidant protein is more obvious in the triplets composed of hydrophobic amino acids, so we infer that the hydrophobicity of the protein can be used to classify the antioxidant proteins.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/MAX-zyx/antioxidant_dataset.git.

AUTHOR CONTRIBUTIONS

YMZ conceived and designed the project. YXZ and YC conducted the experiments and analyzed the data. YXZ and YMZ wrote the manuscript. ZXT and YMZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant 61971119 and Grant 61901103), and the Natural Science Foundation of Heilongjiang Province (Grant LH2019F002).

REFERENCES

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2018). iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* 16, 412–420. doi: 10.1016/j.csbj.2018.10.007

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* 40, 1276–1314. doi: 10.1002/med.21658

Birben, E., Sahiner, U. M., Sackesen, C., Erzurum, S., and Kalayci, O. (2012). Oxidative stress and antioxidant defense. *World Allergy Organ. J.* 5, 9–19.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intellig. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chao, L., Wei, L., and Zou, Q. (2019). SecProMTB: a SVM-based classifier for secretory proteins of *Mycobacterium tuberculosis* with imbalanced data set. *Proteomics* 19:e1900007.

Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem.* 30, 163–172. doi: 10.1002/jcc.21053

Chen, K., Kurgan, L., and Rahbari, M. (2007a). Prediction of protein crystallization using collocation of amino acid pairs. *Biochem. Biophys. Res. Commun.* 355, 764–769. doi: 10.1016/j.bbrc.2007.02.040

Chen, K., Kurgan, L. A., and Ruan, J. (2007b). Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct. Biol.* 7:25. doi: 10.1186/1472-6807-7-25

Chen, K., Kurgan, L. A., and Ruan, J. (2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem.* 29, 1596–1604. doi: 10.1002/jcc.20918

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144.

Fdez-Glez, J., Ruano-Ordas, D., Mendez, J. R., Fdez-Riverola, F., Laza, R., and Pavon, R. (2018). Determining the influence of class imbalance for the triage of biomedical documents. *Curr. Bioinform.* 13, 592–605. doi: 10.2174/1574893612666170718151238

Feng, P., Chen, W., and Lin, H. (2016). Identifying antioxidant proteins by using optimal dipeptide compositions. *Interdiscipl. Sci. Comput. Life Sci.* 8, 186–191. doi: 10.1007/s12539-015-0124-9

Feng, P.-M., Lin, H., and Chen, W. (2013). Identification of antioxidants from sequence information using naive Bayes. *Comput. Math. Methods Med.* 2013:567529.

Guo, M., and Zou, Q. (2019). Perspectives of bioinformatics in big data era. *Curr. Genom.* 20, 79–80. doi: 10.2174/1389202920020190422120915

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorat. Newslett.* 11, 10–18. doi: 10.1145/1656274.1656278

Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S., and Kurata, H. (2020). i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.* 103, 225–234. doi: 10.1007/s11103-020-00988-y

Huo, Y., Xin, L., Kang, C., Wang, M., Ma, Q., and Yu, B. (2020). SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso. *J. Theor. Biol.* 486:110098. doi: 10.1016/j.jtbi.2019.110098

Jiang, Q. H., Wang, G. H., Jin, S. L., Li, Y., and Wang, Y. D. (2013). Predicting human microRNA-disease associations based on support vector machine. *Intern. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/ijdmb.2013.056078

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 559–563.

Li, S. H., Zhang, J., Zhao, Y. W., Dao, F. Y., Ding, H., Chen, W., et al. (2019). iPhoPred: a predictor for identifying phosphorylation sites in human protein. *IEEE Access.* 7, 177517–177528. doi: 10.1109/access.2019.2953951

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22.

Liguori, I., Russo, G., Curcio, F., Bulli, G., Aran, L., Della-Morte, D., et al. (2018). Oxidative stress, aging, and diseases. *Clin. Interv. Aging* 13:757.

Lin, C., Chen, W., Qiu, C., Wu, Y., Krishnan, S., and Zou, Q. (2014). LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* 123, 424–435. doi: 10.1016/j.neucom.2013.08.004

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Chen, S., Yan, K., and Weng, F. (2019a). iRO-PsekGCC: identify DNA replication origins based on Pseudo k-tuple GC Composition. 10:842

Liu, B., Gao, X., and Zhang, H. (2019b). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740

Liu, B., Zhu, Y., and Yan, K. (2019c). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Brief. Bioinform.* 2019:bbz139. doi: 10.1093/bib/bbz139

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, M. L., Su, W., Guan, Z. X., Zhang, D., Chen, W., Liu, L., et al. (2020). An overview on predicting protein subchloroplast localization by using machine learning methods. *Curr. Protein Pept. Sci.* doi: 10.2174/1389203721666200117153412 [E-pub Ahead of Print].

Lv, H., Dao, F. Y., Zhang, D., Guan, Z. X., Yang, H., Su, W., et al. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 23:100991. doi: 10.1016/j.isci.2020.100991

Lv, Z. B., Zhang, J., Ding, H., and Zou, Q. (2020). RF-PseU: a random forest predictor for RNA pseudouridine sites. *Front. Bioeng. Biotechnol.* 8:134. doi: 10.3389/fbioe.2020.00134

Mahmood, T., and Yang, P.-C. (2012). Western blot: technique, theory, and trouble shooting. *N. Am. J. Med. Sci.* 4:429. doi: 10.4103/1947-2714.100998

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019a). AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.* 17, 972–981. doi: 10.1016/j.csbj.2019.06.024

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019c). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O., and Lee, G. (2018a). iBCE-EL: a new ensemble learning framework for improved linear B-Cell epitope prediction. *Front. Immunol.* 9:1695. doi: 10.3389/fimmu.2018.01695

Manavalan, B., Shin, T. H., Kim, M. O., and Lee, G. (2018b). PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* 9:1783. doi: 10.3389/fimmu.2018.01783

McGuffin, L. J., Bryson, K., and Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405. doi: 10.1093/bioinformatics/16.4.404

Meng, C., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Pisoschi, A. M., and Pop, A. (2015). The role of antioxidants in the chemistry of oxidative stress: a review. *Eur. J. Med. Chem.* 97, 55–74. doi: 10.1016/j.ejmech.2015.04.040

Quan, Z., Dariusz, M., Qin, M., and Yungang, X. (2017). scalable data mining algorithms in computational biology and biomedicine. *Biomed. Res. Intern.* 2017:5652041.

Schieber, M., and Chandel, N. S. (2014). ROS function in redox signaling and oxidative stress. *Curr. Biol.* 24, R453–R462.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acade. Sci. U.S.A.* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480.

Tang, H., Chen, W., and Lin, H. (2016). Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol. Biosyst.* 12, 1269–1275. doi: 10.1039/c5mb00883b

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Intern. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformaitcs* 2020:btaa667. doi: 10.1093/bioinformatics/btaa667

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17, 17–18.

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151.

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9(Suppl. 2):S22. doi: 10.1186/1472-6807-7-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. *PLoS One* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, J., Chen, S., Dong, L., and Wang, G. (2020). CHTKC: a robust and efficient k-mer counting algorithm based on a lock-free chaining hash table. *Brief. Bioinform.* 2020:bbaa063.

Wang, M., Yue, L., Cui, X., Chen, C., Zhou, H., Ma, Q., et al. (2020). Prediction of extracellular matrix proteins by fusing multiple feature information, elastic net, and random forest algorithm. *Mathematics* 8:169. doi: 10.3390/math8020169

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35, 2395–2402. doi: 10.1093/bioinformatics/bty995

Xu, L., Liang, G., Shi, S., and Liao, C. (2018). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Intern. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* 45, 12100–12112. doi: 10.1093/nar/gkx870

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yu, L., and Gao, L. (2019). Human pathway-based disease network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1240–1249. doi: 10.1109/tcbb.2017.2774802

Yu, L., Gao, L., and Li, K. (2010). A method based on local density and random walks for complexes detection in protein interaction networks. *J. Bioinform. Comput. Biol.* 8, 47–62. doi: 10.1142/s0219720010005191

Yu, L., Xu, F., and Gao, L. (2020). Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression. *Front. Bioeng. Biotechnol.* 8:8. doi: 10.3389/fbioe.2020.00008

Yu, L., Yao, S. Y., Gao, L., and Zha, Y. H. (2019). Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. *Front. Genet.* 9:745. doi: 10.3389/fgene.2018.00745

Zhang, L., Zhang, C., Gao, R., Yang, R., and Song, Q. (2016). Sequence based prediction of antioxidant proteins using a classifier selection strategy. *PLoS One* 11:e0163274. doi: 10.1371/journal.pone.00163274

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinform.* 21:43. doi: 10.1186/1472-6807-7-43

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017:7049406.

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed. Res. Int.* 2015:861402.

Zhou, S., Zhang, F., and Zhang, L. (2018). Editorial: bioinformatics in biological big data era. *Curr. Bioinform.* 13, 435–436. doi: 10.2174/1574893613051808061231O2

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Chen, L., Huang, T., Zhang, Z., and Xu, Y. (2017). Machine learning and graph analytics in computational biomedicine. *Artif. Intell. Med.* 83:1. doi: 10.1016/j.artmed.2017.09.003

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.