frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fbioe.2020.584807

Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction

Zifan Guo[1], *Pingping Wang*[2], *Zhendong Liu*[3]* *and Yuming Zhao*[4]*

[1] School of Aeronautics and Astronautic, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [2] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China, [3] School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China, [4] Information and Computer Engineering College, Northeast Forestry University, Harbin, China

Thermophilicity is a very important property of proteins, as it sometimes determines denaturation and cell death. Thus, methods for predicting thermophilic proteins and non-thermophilic proteins are of interest and can contribute to the design and engineering of proteins. In this article, we describe the use of feature dimension reduction technology and LIBSVM to identify thermophilic proteins. The highest accuracy obtained by cross-validation was 96.02% with 119 parameters. When using only 16 features, we obtained an accuracy of 93.33%. We discuss the importance of the different characteristics in identification and report a comparison of the performance of support vector machine to that of other methods.

Keywords: support vector machine, thermophilic proteins, feature dimension reduction, amino acid, feature selection

OPEN ACCESS

Edited by:
Xue Xu,
Harvard Medical School,
United States

Reviewed by:
Leyi Wei,
Shandong University, China
Xiangxiang Zeng,
Hunan University, China

***Correspondence:**
Zhendong Liu
liuzd2000@126.com
Yuming Zhao
zym@nefu.edu.cn

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *18 July 2020*
Accepted: *11 September 2020*
Published: *22 October 2020*

Citation:
Guo Z, Wang P, Liu Z and Zhao Y (2020) Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. Front. Bioeng. Biotechnol. 8:584807. doi: 10.3389/fbioe.2020.584807

INTRODUCTION

Temperature is a critical condition for life. Proteins are less stable than other macromolecules, and temperature changes can easily lead to protein denaturation, which can lead to cell death (Kumar et al., 2000). Thus, it is important to develop a highly efficient method for predicting protein thermophilicity, which will contribute to the design of stable proteins. The properties of many proteins are related to their thermal stability. Studies have shown that the thermal stability of proteins is influenced by ion number, salt bridge presence, amino acid composition (AAC), dipeptide composition (DPC), and other factors (Sadeghi et al., 2006; Wang H. et al., 2018; Yin et al., 2020). Zhang and Fang (2006), Li et al. (2018), and Wang Y. et al. (2020) found significant differences in the presence of some dipeptides between thermophilic and mesothermal proteins. In addition, Gromiha et al. (1999) found that protein stability was associated with the balance between packing and solubility.

Many studies have been conducted on methods of distinguishing thermophilic proteins from normal-temperature proteins based on protein properties. Liang et al. (2005) proposed an amino acid coupling model with strong statistical ability to distinguish between thermophilic proteins and mesophilic proteins. LogitBoost Classifier and 20 features were used to distinguish thermophilic proteins by Zhang and Fang (2007) which achieved an overall classification accuracy reaching 88.9%. Montanucci et al. (2008) applied support vector machine (SVM) to investigate the

impacts of mutations on the thermal stability of proteins, and with jackknife cross-validation, they achieved a prediction accuracy of 88%. Recently, Lin and Chen (2011) used feature selection technique and SVM with 30 parameters to predict thermotropic proteins, and the overall accuracy reached 93.27%. These methods have achieved good accuracy, but there remains room for improvement in the number of features used and prediction performance.

In this work, we used the data set of Lin and Chen (2011) after eliminating redundancy to distinguish between thermophilic proteins and non-thermophilic proteins. After feature extraction, MRMD2.0 was applied for feature selection and dimension reduction, and LIBSVM was used to obtain the optimal parameters of the model and establish the prediction model. Finally, from the results of cross-validation, both the number of features and the prediction accuracy were improved; the overall prediction accuracy with only 16 features in AAC was increased to 93.33%, and the highest overall accuracy, attained with 119 parameters, reached 96.02%. In addition, we analyzed the importance of features and demonstrated the strong performance of SVM by comparing this method with other methods.

MATERIALS AND METHODS

Data Sets

In this article, we conducted prediction experiments using two groups of data, namely, a group of thermophilic protein data and a group of non-thermophilic protein data. The data sets were collected by Lin and Chen (2011). Generally, thermophilic proteins and non-thermophilic proteins derive from the corresponding biosome, and optimum growth temperature is the key feature used to distinguish thermophilic and non-thermophilic proteins. Therefore, we used 60°C as the minimum optimum growth temperature for thermophilic proteins and 30°C as the maximum optimum growth temperature for non-thermophilic proteins to avoid the problem of protein denaturation. As a result, 136 prokaryotic genomes conforming to the standard were selected, and their protein sequences were obtained from the Universal Protein Resource.

Next, we screened the protein sequences to increase the quality of the data sets. The filtering process employed the following criteria: (1) the sequence must have manual annotation and evaluation; (2) the protein sequence cannot include ambiguous residue; (3) the sequences cannot be fragments of other proteins; and (4) the sequence cannot be deduced from prediction or homology. After the above screening process, we obtained a total of 1,250 non-thermophilic proteins and 1,329 thermophilic proteins. Next, highly similar sequences were removed by employing the CD-HIT program, resulting in 793 non-thermophilic proteins and 915 thermophilic proteins.

Feature Extraction

Before protein prediction, the features of the protein sequences were extracted to construct the feature vectors (**Figure 1**). For this purpose, iFeature was used, which is a utility toolkit based on python to obtain miscellaneous numerical feature representation schemes for protein sequences (Chen et al., 2018). When using iFeature, users can combine various feature clustering, feature selection, and dimension reduction algorithms to promote the analysis of feature importance and model training. iFeature has been widely tested to ensure the validity of our calculations to further ensure the strength of our work.

We used iFeature to extract the features of the protein sequences from our data set, including AAC (Bhasin and Raghava, 2004; Pan et al., 2018; Chen et al., 2019b; Liu et al., 2019; Shen et al., 2019b; Tang et al., 2019; Li Y. H. et al., 2020), C/T/D composition (CTDC), C/T/D transition (CTDT), conjoint triad (CTriad), dipeptide deviation from the expected mean (DDE) (Saravanan and Gautham, 2015), DPC (Saravanan and Gautham, 2015; Chen et al., 2019a), tripeptide composition (TPC), composition of k-spaced amino acid pairs (CKSAAP), grouped dipeptide composition (GDPC), and grouped tripeptide composition (GTPC). The following is a concise explanation of the feature extraction protocol. In all of the following formulas, n denotes the length of the protein sequence.

AAC

AAC refers to the frequency of each amino acid in a protein or peptide sequence. There are 20 kinds of naturally occurring amino acids, namely, ACDEFGHIKLMNPQRSTVWY, and their frequencies in a sequence can be calculated by the following formula:

$$f(i) = \frac{n(i)}{n},\ i \in \{A, C, D, E, F, \ldots, W, Y\}$$

where $n(i)$ refers to the number of occurrences of amino acid i.

DPC

DPC refers to the frequency of dipeptide combinations in a protein or peptide sequence, which yields 400 descriptors (Cheng J. H. et al., 2018; Tang et al., 2018). It is defined by the following formula:

$$f(x, y) = \frac{n_{xy}}{n-1},\ x, y \in \{A, C, D, E, F, \ldots, W, Y\}$$

where n_{xy} refers to the number of dipeptides denoted by amino acids x and y.

TPC

TPC refers to the frequency of tripeptide combinations in a protein or peptide sequence, which yields 8,000 descriptors (Tan et al., 2019; Zhu et al., 2019). It is defined by the following formula:

$$f(x, y, z) = \frac{n_{xyz}}{n-2},\ x, y, z \in \{A, C, D, E, F, \ldots, W, Y\}$$

where n_{xyz} refers to the number of tripeptides denoted by amino acid combination x, y, and z.

DDE

The DDE eigenvector is constructed by calculating three parameters: dipeptide composition (D_c), theoretical mean value

FIGURE 1 | Study flowchart. (I) The original protein sequence is input for feature extraction. (II) A feature extraction algorithm is used to obtain feature descriptors of each protein. (III) MRMD2.0 is used to rank the importance of features and select features. (IV) Support vector machine is used for parameter optimization and training model establishment. (V) Three parameters are used to evaluate the performance of the model: sensitivity (SE), specificity (SP), and accuracy (ACC).

(T_m), and theoretical variance (T_v). These three parameters and DDE are calculated as follows:

$$D_c(x,y) = \frac{n_{xy}}{n-1}, \quad x,y \in \{A,C,D,E,F,\ldots,W,Y\}$$

where n_{xy} refers to the number of dipeptides displayed by amino acid combination x and y.

$$T_m(x,y) = \frac{C_x}{C_n} \times \frac{C_y}{C_n}, \quad x,y \in \{A,C,D,E,F,\ldots,W,Y\}$$

where C_x and C_y are the number of codons encoding the first and second amino acids, respectively, in dipeptide "x, y," and C_n is the total number of possible codons remaining after removing the 3 terminated codons.

$$T_v(x,y) = \frac{T_m(x,y)\left(1 - T_m(x,y)\right)}{n-1}, \quad x,y \in \{A,C,D,E,F,\ldots,W,Y\}$$

$$DDE(x,y) = \frac{D_c(x,y) - T_m(x,y)}{\sqrt{T_v(x,y)}}$$

GDPC

The GDPC encoding is a change of the DPC descriptor that includes a total of 25 descriptors, defined as follows:

$$f(x,y) = \frac{n_{xy}}{n-1}, \quad x,y \in \{g1,g2,g3,g4,g5\}$$

where n_{xy} refers to the number of dipeptides denoted by amino acid groups x and y.

GTPC

The GTPC is another change of TPC descriptor, which consists of a total of 125 descriptors and is defined as follows:

$$f(x,y,z) = \frac{n_{xyz}}{n-2}, \quad x,y,z \in \{g1,g2,g3,g4,g5\}$$

where n_{xyz} refers to the number of tripeptides denoted by amino acid combination x, y, and z.

CTD

CTD features represent the structural or physicochemical distribution patterns of amino acids in protein or peptide sequences (Dubchak et al., 1999; Tang et al., 2020). Thirteen types of physicochemical properties were used to calculate these characteristics, including hydrophobicity, standardized van der Waals volume, solvent accessibility, polarity, secondary structure, polarizability, and charge. These descriptors were computed by the following procedures: (1) the amino acid sequences were changed into residues with certain structural or physicochemical properties; (2) according to the main cluster of Tomii and Kanehisa (1996) amino acid index, the 20 amino acids were divided into 3 groups according to 7 physicochemical properties.

CTDC

After all 20 amino acids are divided into three groups, the composition descriptor is composed of 3 values, which are the total percentages of group 1, group 2, and group 3 of the protein sequences. The descriptor is calculated as follows:

$$C(x) = \frac{n(x)}{n}, \quad x \in \{\text{group 1}, \text{group 2}, \text{group 3}\}$$

where $n(x)$ refers to the number of occurrences of amino acid x in the encoded sequence.

CTDT

The transformation descriptor T also contains three values. The transition from group 1 to group 2 is the percentage frequency of a residue from group 1 followed by a residue from group 2 or a residue from group 2 followed by a residue from group 1. Transformations between group 2 and group 3 and between group 3 and group 1 are defined in a similar manner. The transformation descriptor can be calculated as follows:

$$T(x,y) = \frac{n(x,y) + n(y,x)}{n-1}, \quad x,y \in \{(\text{group 1}, \text{group 2}), (\text{group 2}, \text{group 3}), (\text{group 3}, \text{group 1})\}$$

where $n(x,y)$ and $n(y,x)$ refer to the numbers of dipeptides denoted by "x, y" and "y, x," respectively, in the protein sequence.

Feature Selection

Feature selection is an important step in the process of protein classification (**Figure 1**) (Feng et al., 2017; Cheng, 2019; Liu, 2019; Yang W. et al., 2019; Zheng et al., 2019; Wang M. et al., 2020; Yang et al., 2020b; Zhao et al., 2020). MRMD2.0 is a very deep feature selection method, which uses the concept of the PageRank algorithm and is combined with methods such as analysis of variance (Scheffe, 1960), minimal redundancy and maximal relevance (Ding and Peng, 2005), maximal information coefficient, and least absolute shrinkage and selection operator (Xu et al., 2017). As a result, MRMD2.0 integrates seven different feature ranking algorithms with PageRank algorithm and detects optimized dimensionality with forward adding strategy. PageRank algorithm was originally used to attach weight value to each target page: pages with large weight values are displayed in the front, whereas pages with small weight values are displayed in the back. Similarly, MRMD2.0 uses PageRank algorithm and several other feature ranking algorithms to generate a corresponding weight value for each feature to form a ranking of the importance of all features.

In this study, MRMD2.0 was used to select features and reduce the dimension of the obtained features to improve the feature prediction ability. By treating each group of features in the previous step with MRMD2.0, we obtained the combination of features with the highest classification accuracy and the importance ranking of each group of features. Generally, the combination of features with the highest classification accuracy has fewer dimensions, so we refer to this process as feature dimension reduction. Based on the classification performance, we ranked the group of features. After combining the features with good classification performance, we applied MRMD2.0 to select them again. Finally, after comparing the results, we obtained the combination of features with the best classification ability.

In addition, we applied MRMD2.0 to obtain the importance ranking of features. On the rank list, higher-ranked features are more predictive; accordingly, we identified the most important features for the classification of thermophilic proteins and non-thermophilic proteins. The resulting information enhances our knowledge of the properties of proteins and can aid the construction of stable proteins in protein engineering.

LIBSVM

In this study, LIBSVM was used to construct models and make predictions (**Figure 1**). LIBSVM is an effective SVM pattern recognition and regression software package designed by Chih-Jen Lin, a professor at Taiwan University, and has been applied in many fields (Lin et al., 2012; Liu et al., 2012, 2017; Ding et al., 2017; Zeng et al., 2017; Wei et al., 2018, 2019; Xu et al., 2018b,c; Cheng et al., 2019b; Deng et al., 2019; Liang et al., 2019; Shen et al., 2019b,a; Su et al., 2019; Yang H. et al., 2019; Li F. et al., 2020; Wang H. et al., 2020; Yang et al., 2020a; Zhang et al., 2020). Before training SVM on a problem, the parameters must be specified (Jiang et al., 2013; Zhao et al., 2015, 2017). We selected the best parameters, C and g, through a simple tool provided by LIBSVM for evaluating a grid of parameters. The accuracy for each parameter setting is obtained in LIBSVM, allowing the parameters with the highest cross-validation accuracy to be determined. Next, we trained the whole data set with the best parameters C and g to obtain the prediction model. Finally, we tested and predicted our data set with the obtained model.

Performance Measurement

We used three commonly used indicators to evaluate model performance: sensitivity (SE), specificity (SP), and accuracy (ACC) (**Figure 1**) (Wang et al., 2010; Wei et al., 2017a,b; Zhang et al., 2018; Cheng et al., 2019a; Ding et al., 2019a; Junwei et al., 2019; Liang et al., 2019; Liu and Li, 2019; Tian et al., 2019; Jia et al., 2020; Liu and Chen, 2020; Li J. et al., 2020; Lv et al., 2020; Wang Z. et al., 2020). They are described as follows:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

where TN, TP, FN, and FP refer to the numbers of correctly predicted non-thermophilic proteins, correctly predicted non-thermophilic proteins, incorrectly predicted non-thermophilic proteins, and incorrectly predicted thermophilic proteins, respectively. SE and SP indicators measure the predictive ability of a model in positive and negative situations, respectively, and ACC is used to evaluate the overall performance of a prediction model (Wang et al., 2008; Zou et al., 2017a,b; Cheng L. et al., 2018; Wang G. et al., 2018; Xue et al., 2018; Xu et al., 2018a, 2019; Ding et al., 2019b; Shen et al., 2019b; Yang, 2019; Zeng et al., 2019; Fu et al., 2020; Hong et al., 2020).

RESULTS AND DISCUSSION

Identification of Protein Thermostability

The results of feature selection by using MRMD2.0 are shown in **Table 1**. Among them, features with good classification performance include AAC, DPC, CTDC, and dipeptide deviation from the expected mean. However, although the classification ACC of dipeptide deviation from the expected mean after dimension reduction reached 85.6%, it had 365-dimensional features. Considering the excessive dimension and the unexceptional performance, only AAC, DPC, and CTDC were subsequently combined for classification.

Next, based on LIBSVM and grid parameter optimization, we used various combinations of these three features to construct models and perform cross-validation for our data sets. The results are shown in **Table 2**. The overall ACC of three schemes is higher than that of Lin and Chen (2011) (93%).

Initially, we used AAC with 16 dimensions alone to build a prediction model for the data set, achieving an overall ACC rate of 93.33% through cross-validation, which is slightly higher than that of Lin and Chen (2011). In addition, Zhang and Fang (2006) and Gromiha and Suresh (2010) used all 20 amino acids

TABLE 1 | The results of feature selection by using MRMD2.0.

Feature	Dimensions	Accuracy (%)
AAC	16/20	87.94
DPC	103/400	87.00
DDE	365/400	85.60
CTDC	33/39	85.01
CTDT	39/39	80.50
CTriad	338/343	79.80
CKSAAP	143/150	79.04
GTPC	107/125	78.63
GDPC	13/25	78.57
TPC	1,008/1023	77.11

The two numbers in the second column of the table are the number after dimension reduction and the number before dimension reduction.

TABLE 2 | The results of classification using SVM and various feature combinations.

Feature combination	SE (%)	SN (%)	Accuracy (%)
The method of Lin and Chen (2011)	93.77	92.69	93.27
AAC (16)	93.44	93.19	93.33
AAC (16) + CTDC (33)	93.77	92.81	93.33
AAC (16) + DPC (103)	95.85	96.22	96.02

The numbers in parentheses in the first column of the table represent the number of arguments to the feature preceding the parentheses.

TABLE 3 | The results of classification accuracy using LIBSVM and various combinations of important features.

Dimension	Feature	Accuracy (%)
1	K	76.41
2	K + D	77.50
3	K + D + LK	78.29

A plus sign in the second column of the table indicates the use of these characteristics for model training and classification. For example, "K + D" indicates the modeling and classification of the data sets with the two-dimension characteristics K and D.

composition to predict the thermostability of protein, and their overall ACC was 90.5 and 89%, respectively. Furthermore, Wang and Li (2014) enhanced the ACC to 95% by selecting 9 AAC and 38 DPC using a genetic algorithm. In contrast, the scheme used only 16 parameters, but the ACC reached 93.33%, which is fewer than the dimensions used in previous studies. The results show that AAC plays an important role in the identification of thermophilic proteins.

The top two features in **Table 3** were AAC and DPC. The model constructed with 16 parameters of AAC and 103 parameters of DPC achieved the highest overall ACC of 96.02%. The SE and SP of this method were 95.85 and 96.22%, respectively, which indicates that the predictive ability of this model in both positive and negative situations is excellent.

In addition, we used the combination of AAC with 16 dimensions and CTDC with 33 dimensions to build a prediction model and obtained the same overall ACC as the first model. However, this second model had higher SE and lower SP than the first model, indicating that it was slightly inferior to the model built with 16 dimensions of AAC.

Feature Importance

We aimed to identify the most important features of the method with 119 parameters that can achieve the highest ACC and analyze them. To assess feature importance, first, we used MRMD2.0 to rank all 119 features by importance. We found that the top three features were K, D, and LK (Feature K is the percentage of lysine in the amino acid sequence, feature D is the percentage of aspartic acid in the amino acid sequence, and feature LK is the percentage content of the dipeptide consisting of leucine and lysine in the amino acid sequence). These three features are arguably the most predictive among the 119 features for the classification of thermophilic proteins.

Next, to obtain the classification performance of the above features, we used one-dimensional (K), two-dimensional (K and D), and three-dimensional (K, D, and LK) features to classify our data set based on LIBSVM. The results are shown in **Table 3**.

As seen from **Table 3**, the classification ACC of the K feature alone reached 76.41%, whereas the ACC achieved with K combined with D and LK was only slightly greater. To better analyze the classification ability of these three important features, we constructed a violin diagram, scatter diagram, and 3D scatter diagram for the 1-, 2-, and 3-dimension features. The results are shown in **Figure 2**.

As seen from **Figure 2A**, the K value of the thermophilic proteome is concentrated ∼0.08, whereas the K value of the non-thermophilic proteome is concentrated ∼0.03. These results indicate that the K feature can well distinguish thermophilic proteins from non-thermophilic proteins, a finding of great significance for the identification of the thermophilic properties of proteins. All three panels reveal obvious differences in the distribution pattern between the two data sets, which indicates that these features have strong recognition ability and good performance in distinguishing thermophilic proteins from non-thermophilic proteins, as shown in **Table 3**.

Comparison With Other Classification Methods

To reveal the advantage of our method, we applied six other classification methods to train our data sets based on the Waikato environment for knowledge analysis (Weka) tool (Witten and Frank, 2002): logistic, random forest, BayesNet, logistic model trees (LMTs), J48, and reduced error pruning tree (REPTree).

We used the combination with the highest overall ACC in this article (16 features in AAC and 103 features in DPC) as the input, and we used the above classifiers to predict the data set to obtain the SE, SP, and ACC of each method. To ensure a robust comparison, we also used cross-validation to predict the data set. By comparing the performance of different methods, the performance of

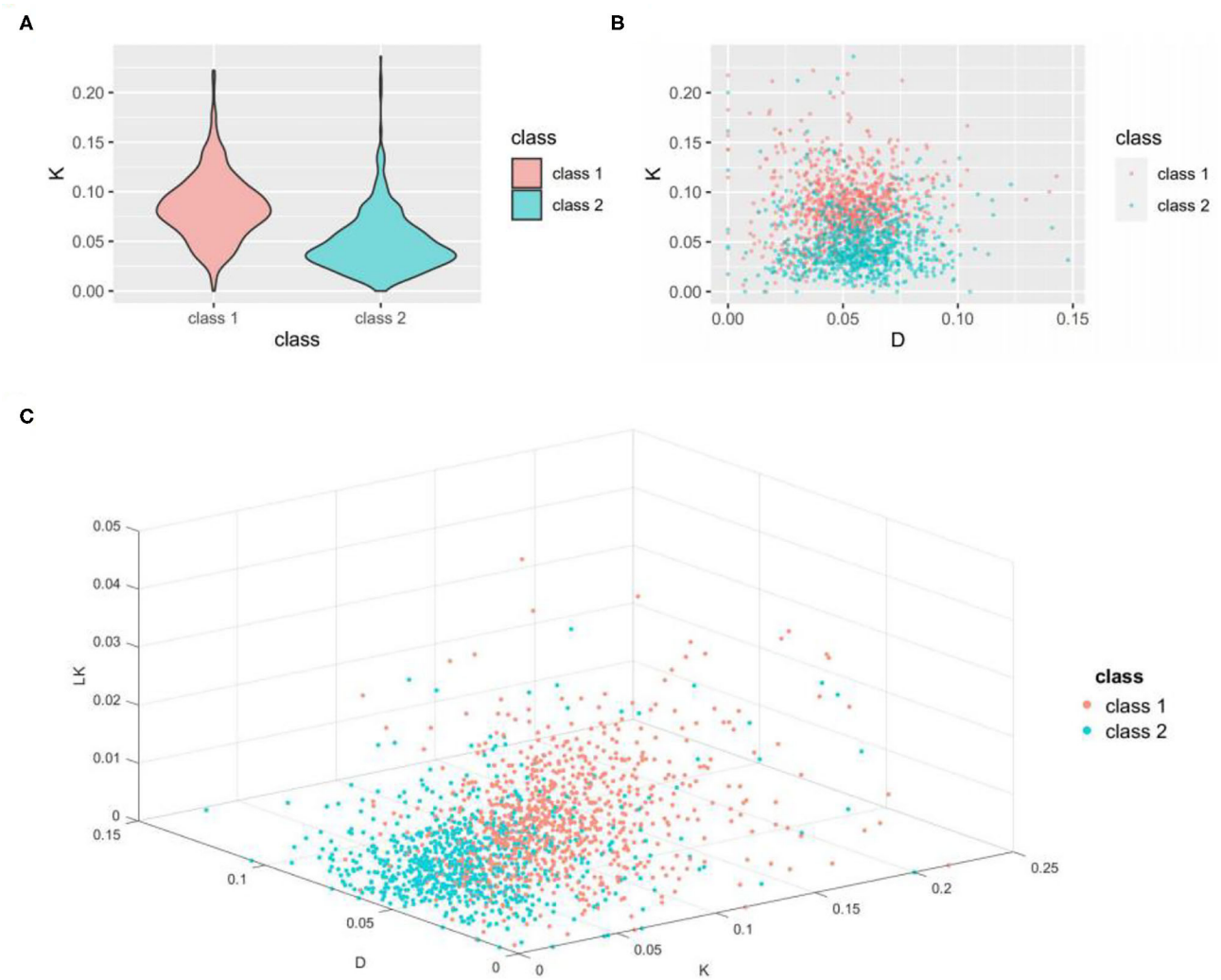

FIGURE 2 | Visualization of the ability of important features to classify thermophilic and non-thermophilic proteins. **(A)** is a violin diagram of the K feature. **(B)** is a scatter diagram of the K feature and D feature. **(C)** is a 3D scatter diagram of the K, D, and LK features. K is the percentage of lysine in the amino acid sequence, D is the percentage of aspartic acid in the amino acid sequence, and LK is the percentage content of the dipeptide consisting of leucine and lysine in the amino acid sequence.

TABLE 4 | The performance of different classification methods in the prediction of the data sets.

Classification method	SE (%)	SN (%)	Accuracy (%)
SVM (this article)	95.85	96.22	96.02
LMT	92.35	90.29	91.40
Logistic	91.15	88.90	90.11
Random Forest	91.69	87.51	89.75
BayesNet	88.08	86.25	87.24
REPTree	83.60	84.62	84.07
J48	83.50	80.33	82.03

different classifiers was evaluated. The prediction results of each method applied to the data set are shown in **Table 4**.

It can be seen from **Table 4** that the SVM we used in this study achieved the best performance; the SE, SP, and ACC of the other methods were all lower than those of the SVM method of this article. To visualize the data, we constructed a cluster histogram of the performance of the different methods, shown in **Figure 3**.

The advantage of using SVM to predict data sets is apparent from the histogram.

CONCLUSION

In this article, we distinguished 915 thermophilic proteins and 793 non-thermophilic proteins. We applied iFeature to extract the features of the protein sequences. MRMD2.0 was used to reduce the dimensions of features and select the ones that performed the best. LIBSVM was used to optimize the parameters and establish the prediction model. As a result, the overall ACC

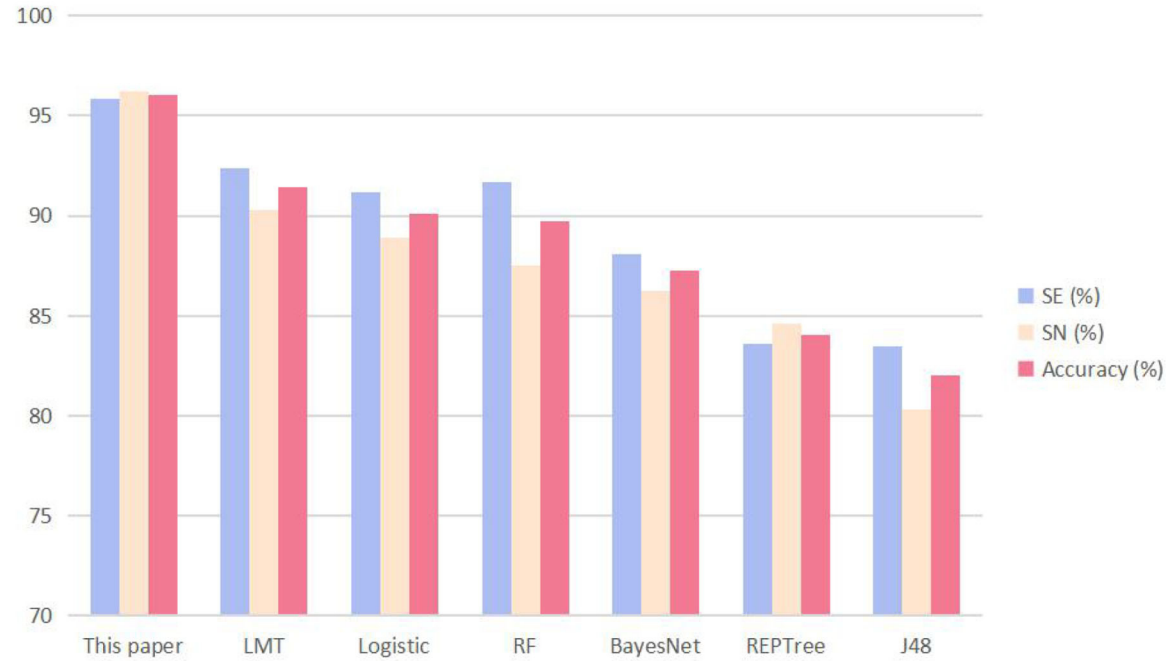

FIGURE 3 | The performance of the method described in this article and other six predictors when the input is 16 parameters of amino acid composition and 103 parameters of dipeptide composition. The performance metrics are sensitivity (SE), specificity (SP), and accuracy (ACC).

was improved, which reached 96.02% under cross-validation. Furthermore, we constructed a prediction model by LIBSVM with 16 parameters, and the ACC determined by cross-validation was 93.33%. In addition, we found that the K feature played a significant role in the identification. Finally, we demonstrated the advantage of SVM by comparing its performance with that of other methods. We aim to analyze information, such as the family of misclassified proteins, to optimize our method in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: doi: 10.1016/j.mimet.2010.10.013.

AUTHOR CONTRIBUTIONS

ZG made the design of the subject and the whole idea of the whole experiment, did comparative experiments, and the analysis of the experiment. PW did experimental data analysis. ZL and YZ analyzed the results of the experiment and made some improvements to this paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of China (Nos. 61971119 and 61672328).

REFERENCES

Bhasin, M., and Raghava, G. P. S. (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279, 23262–23266. doi: 10.1074/jbc.M401932200

Chen, W., Feng, P., Liu, T., and Jin, D. (2019b). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.* 20, 224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Feng, P., and Nie, F. (2019a). *iATP*: a sequence based method for identifying anti-tubercular peptides. *Med. Chem.* 16, 620–625. doi: 10.2174/1573406415666191002152441

Chen, Z., Zhao, P., Li, F., Leier, A., Marquezlago, T. T., Wang, Y., et al. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cheng, J. H., Yang, H., Liu, M. L., Su, W., Feng, P. M., Ding, H., et al. (2018). Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab.* 180, 64–69. doi: 10.1016/j.chemolab.2018.07.006

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19(Suppl. 1):919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019a). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J., Tan, R., et al. (2019b). Exposing the causal effect of body mass index on the risk of type 2

Diabetes mellitus: a mendelian randomization study. *Front. Genet.* 10:94. doi: 10.3389/fgene.2019.00094

Deng, L., Wang, J., and Zhang, J. (2019). Predicting gene ontology function of human MicroRNAs by integrating multiple networks. *Front. Genet.* 10:3. doi: 10.3389/fgene.2019.00003

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/S0219720005001004

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inf. Sci.* 418–419, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019a). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Ding, Y., Tang, J., and Guo, F. (2019b). Identification of drug-side effect association via semi-supervised model and multiple kernel learning. *IEEE J. Biomed. Health Inform.* 23, 2619–2632. doi: 10.1109/JBHI.2018.2883834

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S. (1999). Recognition of a protein fold in the context of the SCOP classification. *Proteins* 35, 401–407.

Feng, P., Ding, H., Lin, H., and Chen, W. (2017). AOD: the antioxidant protein database. *Sci. Rep.* 7:7449. doi: 10.1038/s41598-017-08115-6

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Gromiha, M. M., Oobatake, M., and Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem.* 82, 51–67. doi: 10.1016/S0301-4622(99)00103-9

Gromiha, M. M., and Suresh, M. X. (2010). Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins* 70, 1274–1279. doi: 10.1002/prot.21616

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043. doi: 10.1093/bioinformatics/btz694

Jia, C., Bi, Y., Chen, J., Leier, A., Li, F., and Song, J. (2020). PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. *Bioinformatics* 36, 4276–4282. doi: 10.1093/bioinformatics/btaa522

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. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/IJDMB.2013.056078

Junwei, H., Xudong, H., Qingfei, K., and Liang, C. (2019). psSubpathway: a software package for flexible identification of phenotype-specific subpathways in cancer progression. *Bioinformatics* 36, 2303–2305. doi: 10.1093/bioinformatics/btz894

Kumar, S., Tsai, C., and Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein Eng.* 13, 179–191. doi: 10.1093/protein/13.3.179

Li, F., Zhou, Y., Zhang, X., Tang, J., Yang, Q., Zhang, Y., et al. (2020). SSizer: determining the sample sufficiency for comparative biological study. *J. Mol. Biol.* 432, 3411–3421. doi: 10.1016/j.jmb.2020.01.027

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2020.2977091. [Epub ahead of print].

Li, Y. H., Li, X. X., Hong, J. J., Wang, Y. X., Fu, J. B., Yang, H., et al. (2020). Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief. Bioinform.* 21, 649–662. doi: 10.1093/bib/bby130

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* 46, D1121–D1127. doi: 10.1093/nar/gkx1076

Liang, C., Changlu, Q., He, Z., Tongze, F., and Xue, Z. (2019). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D554–D560. doi: 10.1093/nar/gkz843

Liang, H., Huang, C., Ko, M., and Hwang, J. (2005). Amino acid coupling patterns in thermophilic proteins. *Proteins* 59, 58–63. doi: 10.1002/prot.20386

Lin, H., and Chen, W. (2011). Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods* 84, 67–70. doi: 10.1016/j.mimet.2010.10.013

Lin, H., Ding, C., Song, Q., Yang, P., Ding, H., Deng, K. J., et al. (2012). The prediction of protein structural class using averaged chemical shifts. *J. Biomol. Struct. Dyn.* 29, 643–649. doi: 10.1080/07391102.2011.672628

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., Gao, X., and Zhang, H. (2019). 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., Li, C., and Yan, K. (2012). DeepSVM-fold: protein fold recognition by combining support vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* 21, 1733–1741. doi: 10.1093/bib/bbz098

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, K., and Chen, W. (2020). iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics* 36, 3336–3342. doi: 10.1093/bioinformatics/btaa155

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 905–915. doi: 10.1109/TCBB.2016.2550432

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

Montanucci, L., Fariselli, P., Martelli, P. L., and Casadio, R. (2008). Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics* 2008, 190–195. doi: 10.1093/bioinformatics/btn166

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 34, 1473–1480. doi: 10.1093/bioinformatics/btx822

Sadeghi, M., Naderimanesh, H., Zarrabi, M., and Ranjbar, B. (2006). Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.* 119, 256–270. doi: 10.1016/j.bpc.2005.09.018

Saravanan, V., and Gautham, N. (2015). Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. *OMICS* 19, 648–658. doi: 10.1089/omi.2015.0095

Scheffe, H. (1960). The analysis of variance. *Soil Sci.* 89:360. doi: 10.1097/00010694-196006000-00016

Shen, C., Jiang, L., Ding, Y., Tang, J., and Guo, F. (2019b). LPI-KTASLP. prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/ACCESS.2019.2894225

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2019a). Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform.* 21, 1628–1640. doi: 10.1093/bib/bbz106

Shen, Y., Tang, J., and Guo, F. (2019b). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J. Theor. Biol.* 462, 230–239. doi: 10.1016/j.jtbi.2018.11.012

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

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. doi: 10.3934/mbe.2019123

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. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2020). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief. Bioinform.* 21, 621–636. doi: 10.1093/bib/bby127

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome

quantification by optimizing data manipulation chains. *Mol. Cell. Proteomics* 18, 1683–1699. doi: 10.1074/mcp.RA118.001169

Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q., and Yu, B. (2019). Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. *J. Theor. Biol.* 462, 329–346. doi: 10.1016/j.jtbi.2018.11.011

Tomii, K., and Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* 9, 27–36. doi: 10.1093/protein/9.1.27

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. doi: 10.1093/nar/gkx1096

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: 10.1186/1471-2164-9-S2-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, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert-schmidt independence criterion. *Neurocomputing* 383, 257–269. doi: 10.1016/j.neucom.2019.11.103

Wang, H., Liu, C., and Deng, L. (2018). Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep*. 8:14285. doi: 10.1038/s41598-018-32511-1

Wang, L. Q., and Li, C. F. (2014). Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification. *Biotechnol. Lett*. 36, 1963–1969. doi: 10.1007/s10529-014-1577-3

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, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., et al. (2020). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res*. 48, D1031–D1041. doi: 10.1093/nar/gkz981

Wang, Z., He, W., Tang, J., and Guo, F. (2020). Identification of highest-affinity binding sites of yeast transcription factor families. *J. Chem. Inf. Model*. 60, 1876–1883. doi: 10.1021/acs.jcim.9b01012

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med*. 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Trans. Comp. Biol. Bioinform*. 16, 1264–1273. doi: 10.1109/TCBB.2017.2670558

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med*. 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Witten, I. H., and Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *SIGMOD Rec.* 31, 76–77. doi: 10.1145/507338.507355

Witten, I. H., and Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *SIGMOD Rec.* 31, 76-77. doi: 10.1145/507338.507355

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang C-C. (2019). k-skip-n-gram-RF: a random forest based method for Alzheimer's disease protein identification . *Front. Genet*. 10:33. doi: 10.3389/fgene.2019.00033

Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2018a). An efficient classifier for alzheimer's disease genes identification. *Molecules* 23:3140. doi: 10.3390/molecules23123140

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

Xu, L., Liang, G., Wang, L., and Liao, C. (2018c). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:158. doi: 10.3390/genes9030158

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

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci*. 9, 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yang, C. (2019). Interaction of cell and gene therapy with the immune system. *Curr. Gene Ther*. 19, 69–70. doi: 10.2174/1566523219019072211294

Yang, H., Yang, W., Dao, F. Y., Lv, H., Ding, H., Chen, W., et al. (2019). A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinform*. 21, 1568–1580. doi: 10.1093/bib/bbz123

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2020a). Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief. Bioinform*. 21, 1058–1068. doi: 10.1093/bib/bbz049

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., et al. (2020b). NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res*. 48, W436–W448. doi: 10.1093/nar/gkaa258

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

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2020). VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res*. 48, D1042–D1050. doi: 10.1093/nar/gkz779

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, F., Ma, A., Wang, Z., Ma, Q., Liu, B., Huang, L., et al. (2018). A central edge selection based overlapping community detection algorithm for the detection of overlapping structures in protein–protein interaction *networks. Molecules* 23:2633. doi: 10.3390/molecules23102633

Zhang, G., and Fang, B. (2006). Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. *Process Biochem*. 41, 552–556. doi: 10.1016/j.procbio.2005.09.003

Zhang, G., and Fang, B. (2007). LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J. Biotechnol*. 127, 417–424. doi: 10.1016/j.jbiotec.2006.07.020

Zhang, Z. Y., Yang, Y. H., Ding, H., Wang, D., Chen, W., and Lin, H. (2020). Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform*. doi: 10.1093/bib/bbz177. [Epub ahead of print].

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/s12859-020-3388-y

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. doi: 10.1155/2017/7049406

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed. Res. Int*. 2015:861402. doi: 10.1155/2015/861402

Zheng, N., Wang, K., Zhan, W., and Deng, L. (2019). Targeting virus-host protein interactions: feature extraction and machine learning approaches.

Curr. Drug Metab. 20, 177–184. doi: 10.2174/1389200219666180829121038

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. (2017a). Machine learning and graph analytics in computational biomedicine. Artificial intelligence in medicine. *Artif. Intell. Med.* 83:1. doi: 10.1016/j.artmed.2017.09.003

Zou, Q., Mrozek, D., Ma, Q., and Xu, Y. (2017b). Scalable data mining algorithms in computational biology and biomedicine. *Biomed Res. Int*. 2017:5652041. doi: 10.1155/2017/5652041

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.