Hindawi
BioMed Research International
Volume 2020, Article ID 9235920, 11 pages
https://doi.org/10.1155/2020/9235920

Research Article

Identification of Human Enzymes Using Amino Acid Composition and the Composition of *k*-Spaced Amino Acid Pairs

Lifu Zhang,[1,2] **Benzhi Dong**,[3] **Zhixia Teng**,[3] **Ying Zhang**,[4] **and Liran Juan**[5]

[1]*School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu, China*
[2]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China*
[3]*Information and Computer Engineering College, Northeast Forestry University, Harbin, China*
[4]*Department of Pharmacy, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China*
[5]*School of Life Science and Technology, Harbin Institute of Technology, Harbin, China*

Correspondence should be addressed to Ying Zhang; zhangying_hmu@163.com and Liran Juan; lrjuan@hit.edu.cn

Received 14 March 2020; Accepted 22 April 2020; Published 27 May 2020

Guest Editor: Qin Ma

Enzymes are proteins that can efficiently catalyze specific biochemical reactions, and they are widely present in the human body. Developing an efficient method to identify human enzymes is vital to select enzymes from the vast number of human proteins and to investigate their functions. Nevertheless, only a limited amount of research has been conducted on the classification of human enzymes and nonenzymes. In this work, we developed a support vector machine- (SVM-) based predictor to classify human enzymes using the amino acid composition (AAC), the composition of *k*-spaced amino acid pairs (CKSAAP), and selected informative amino acid pairs through the use of a feature selection technique. A training dataset including 1117 human enzymes and 2099 nonenzymes and a test dataset including 684 human enzymes and 1270 nonenzymes were constructed to train and test the proposed model. The results of jackknife cross-validation showed that the overall accuracy was 76.46% for the training set and 76.21% for the test set, which are higher than the 72.6% achieved in previous research. Furthermore, various feature extraction methods and mainstream classifiers were compared in this task, and informative feature parameters of *k*-spaced amino acid pairs were selected and compared. The results suggest that our classifier can be used in human enzyme identification effectively and efficiently and can help to understand their functions and develop new drugs.

1. Introduction

Enzymes, also known as biocatalysts, are proteins that can catalyze chemical reactions in living cells efficiently and specifically, and they play a key role in the survival of humans, other animals, and plants. Over the last few decades, enzymes in increasing numbers have been identified and have been found to have a variety of properties and play diverse roles in the survival, growth, and development of organisms.

Depending on the properties of the reaction catalyzed, enzymes are classified into six classes according to enzyme commission (EC) numbers [1]: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. Owing to the specificity of enzymes, i.e., an enzyme can only catalyze a specific chemical reaction in a cell, accurately classifying and predicting enzyme classes is of vital importance when searching for unknown enzymes and developing new drugs, including zymin.

The traditional approach to the identification of proteins through wet experimental methods has typically been time and resource intensive. With the development of protein sequencing technology and improvements in computing power, computational methods based on amino acid sequence data of peptides, especially machine learning methods, have been widely used to classify and predict the function of diverse classes of proteins [2–7].

Currently, several researchers have focused on developing methods that can be used for the identification of enzymes. Jensen et al. first predicted enzyme classes using sequence-based physicochemical features and an Artificial Neural Network (ANN) in 2002 [8]. Chou and Cai proposed the GO-PseAAC predictor, which combined gene ontology

(GO) and Pseudo amino acid composition (PseAAC) as features to search for and used the nearest neighbor algorithm approach [9]. Later, Cai et al. first applied the SVM algorithm to enzyme classification [10] and combined functional domain composition (FunD) with PseAAC to predict the classes of enzymes [11, 12]. Furthermore, a predictor named EzyPred was developed by Shen and Chou that uses FunD and the Pseudo position-specific scoring matrix (PsePSSM) as features [13]. In 2009, Nasibov and Kandemir-Cavas classified enzymes by the K-nearest neighbor (KNN) method and the minimum distance-based predictor using AAC [14]. Concu et al. provided a distinctive method using the 3D structure rather than sequence information [15]. Qiu et al. developed a method based on PseAAC and discrete wavelet transform (WT) that was trained by the SVM algorithm [16]. Shi and Hu used low-frequency power spectral density and increment of diversity, combined with AAC and PseAAC, and built an SVM-based predictor [17]. In addition, Zou et al. introduced a multilabel learning method to identify multifunctional enzymes [18]. Later, a new method was put forward by Niu et al. that used a protein-protein network [19]. In recent years, deep-learning methods like convolutional neural networks were used for the classification of enzymes and achieved good results [20, 21].

All of these classification methods improved the classification performance based on previous research. Nevertheless, all of these researchers concentrated on classifying different types of enzymes, and very few methods have been developed to predict whether a protein is an enzyme or a nonenzyme. Wu et al. devoted themselves to this issue and designed an SVM-based method combining PseAAC with the rigidity [22], flexibility, and irreplaceability of amino acids to identify human enzyme classes. However, this method only reached an overall accuracy of 72.6% by 5-fold cross-validation using 372 features, and thus, the performance of this task needs to be further improved.

On the basis of the above research, in this work, we developed a new machine learning method to classify human enzymes and nonenzymes. First, we introduced a feature representation strategy based on AAC and the composition of k-spaced amino acid pairs (CKSAAP). Next, for features represented by the methods above, the feature selection technique based on analyses of variance (ANOVA) was applied to minimize the features we used and to improve its overall accuracy. Finally, the selected features were fed into the classifiers found from SVM for training. As a result, an accuracy of 76.46% and 76.21% by 6-fold cross-validation was achieved in the training set and test set, respectively, by using 40 feature parameters. Furthermore, the performances of different feature representation strategies under the SVM classifier and the performances of different classifiers were compared and discussed, and important feature parameters in this task were selected and compared.

2. Materials and Methods

2.1. Datasets. The training sequence data used in this study were first reported by Wu et al. [22] and were obtained from the Universal Protein Resource (UniProt), the protein database with the most abundant information and resources; the training sequence data were composed of data from three databases: Swiss-Prot, TrEMBL, and PIR-PSD [23]. Six subclasses of human enzymes and nonhuman enzymes can be filtered and downloaded for free. To ensure the correctness and representativeness of the training data, the following data preprocessing process was used: (1) Human enzyme sequences of enzymes whose function had not been experimentally verified and those labeled as fragments were eliminated. (2) Enzyme sequences containing ambiguous residues ("B," "J," "O," "U," "X," and "Z") were excluded. (3) The CD-HIT program was applied to remove highly similar enzyme sequences using 30% as the cutoff of sequence identity [24, 25].

After the above data preprocessing steps were completed, 1117 human enzymes and 2099 nonhuman enzymes were selected as training sequences in the analysis. Among them, the human enzyme sequences consist of 6 subclasses, as shown in Figure 1(a), with the overall workflow in our study shown in Figure 1(b).

Furthermore, to evaluate the effect of the model more accurately, a set of test data was selected from the dataset used by Cai and Chou [11] and downloaded from UniProt [23]; these data included a total of 1954 sequences, including 684 enzymes and 1270 nonhuman enzyme sequences, respectively.

2.2. Feature Extraction. One of the most important steps in our method was to extract the feature vector of the selected sequences. Many works have focused on feature extraction of proteins. AAC [26, 27], dipeptide composition (DPC) [28, 29], Geary correction [30], composition-transition-distribution [10, 31, 32], PseAAC [33–37], and other feature extraction methods [38–40] have been proposed and widely applied to describe different kinds of protein primary sequences. Here, we presented and then applied AAC and CKSAAP to extract features.

The AAC encoding strategy calculates the frequency of each type of the 20 amino acids in a primary protein sequence [26], which can be formulated as follows:

$$\begin{cases} \mathscr{R}_{\mathrm{AAC}} = [\mathrm{f}(1)\mathrm{f}(2) \cdots \mathrm{f}(i) \cdots \mathrm{f}(20)]_{20}^{T}, \\ f(i) = \dfrac{N(i)}{L} (1 \le i \le 20), \end{cases} \tag{1}$$

where $N(i)$ denotes the number of the amino acid types i (i.e., A, C, D, E, etc.) and L denotes the length of the sequence. This strategy obtains a 20-D feature vector for each primary sequence.

The CKSAAP encoding strategy reflects the short-range interaction of the sequence. The frequency of 400 amino acid pairs in k-space is calculated using this strategy [41]. The frequency can be defined as follows:

$$\begin{cases} \mathscr{R}_{\mathrm{CKSAAP}} = [f(1,1)\, f(1,2) \cdots f(i,j) \cdots f(20,20)]_{400}^{T}, \\ f(i,j) = \dfrac{N(i,j)}{L-k} (1 \le i, j \le 20), \end{cases} \tag{2}$$

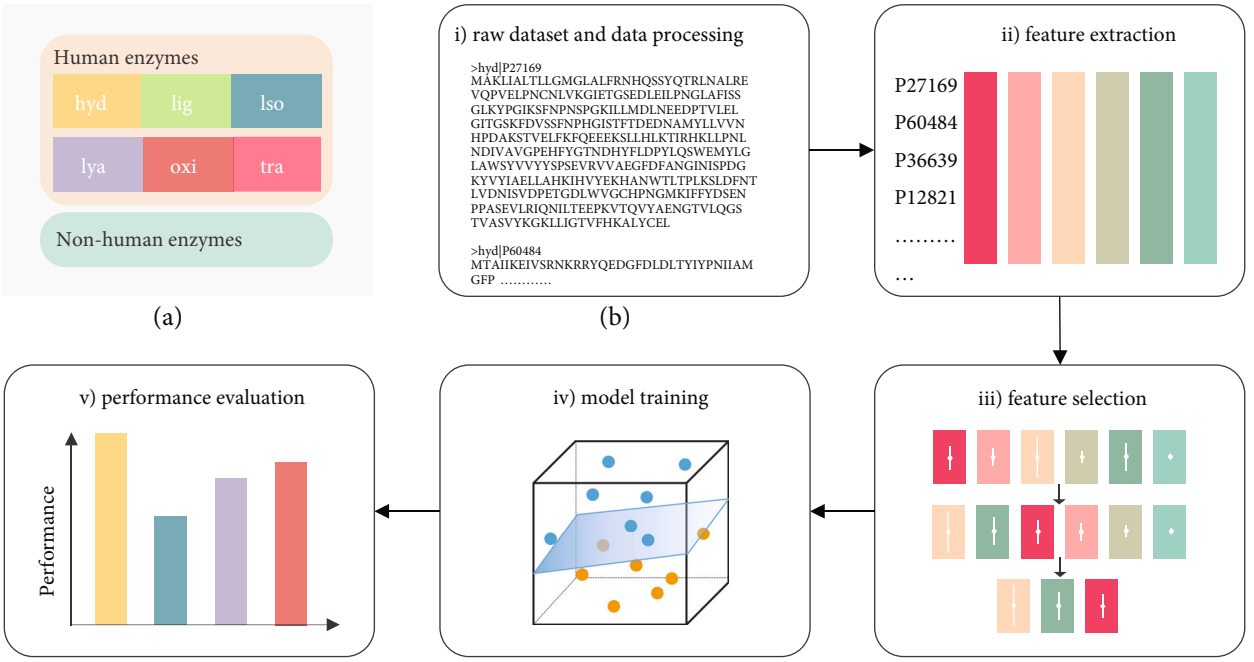

FIGURE 1: Overall workflow. (a) The original sequence dataset used. The dataset consists of human enzymes and nonhuman enzymes. Among them, human enzymes consist of 6 subsets, which represent the catalytic effects on different types of biochemical reactions: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. (b) The workflow of our study. Raw protein sequences were first preprocessed and fed into a feature extraction process, and then, a three-step feature selection technique was used to reduce feature parameters. Last, the selected feature parameters were used to train an SVM-based model, and the performance of the model was evaluated by several evaluation indexes.

where $N(i, j)$ denotes the number of the amino acid types i and j in k-space. L denotes the length of the sequence. This strategy obtains a 400-D feature vector for each primary sequence. Taking $k = 1$ as an example, there are 400 amino acid pairs in 1-space, i.e., A*A, A*C, A*D, etc., where * denotes other amino acids as the gap [42]. In this research, $k = 0$, 1, 2, 3, 4, and 5 are used to extract features and measure the comparative effectiveness. Therefore, the dipeptide composition (DPC) is the same descriptor as CKSAAP when $k = 0$ [43]. Moreover, in our work, features of sequences are extracted by the iFeature toolkit [44].

2.3. Feature Selection. Feature selection was utilized to optimize the prediction model and improve the accuracy of the human-enzyme classification task. In previous research, principal component analysis (PCA), the minimal redundancy maximal relevance (mRMR) algorithm [45, 46], the maximum relevance maximum distance (MRMD) algorithm [47], the genetic algorithm, etc., were proposed for feature selection and applied in protein classification. Here, ANOVA is used to select the most representative features.

ANOVA is an effective method used in statistics to test for a significant relationship between the selected variable and group variables [48, 49]. In our paper, ANOVA can be applied to measure the correlation between a selected feature and all features. The F statistic ($F(\delta)$) of a feature δ is defined as follows:

$$F(\delta) = \frac{s^2_{\mathrm{MSB}}(\delta)}{s^2_{\mathrm{MSW}}(\delta)}, \tag{3}$$

where $s^2_{\mathrm{MSB}}(\delta)$ and $s^2_{\mathrm{MSW}}(\delta)$ represent the mean square between (MSB) and the mean square within (MSW), respectively, which can be interpreted as the sample variance between groups and the sample variance within groups. In the theory of statistics, $F(\delta)$ satisfies the F-distribution, which is used for the significance test. However, in our study, we only focused on the relative values of $F(\delta)$ to indicate the correlation between the feature and the overall size. Features with a larger $F(\delta)$ are selected because a larger $F(\delta)$ implies that they are more strongly related to the group features and more likely to contribute to the classification.

2.4. Support Vector Machine. The SVM algorithm is one of the most popular machine learning algorithms which has been successfully applied in many areas [50–58]. The SVM algorithm is based on statistical learning theory and is widely used in various domains. In the field of protein prediction, SVM has been applied to predicting protein category, secondary structure, physical and chemical properties, etc. and has achieved remarkable results [31, 59–63].

The core idea of SVM is to map the vectors from a low-dimension input space to a high-dimension Hilbert space, in which a linear separating hyperplane is constructed by a kernel function, and to try to maximize the margin among the support vectors of each class by adjusting the linear separating hyperplane. Usually, varieties of kernel functions can be used in SVM algorithms, including linear function, polynomial function, sigmoid function, and radial basis function (RBF). Previous research has shown that RBF performs much better than the other three kinds of kernel functions. Hence, RBF was used in our work as the kernel function [31, 59–63].

Table 1: Accuracy of models trained with various feature parameters added into AAC by 6-fold cross-validation.

Feature parameters added into AAC	Feature selection method	Added number of features/total number of features	Accuracy
CTD-C [10]	mRMR	20/39	75.1547%
CTriad [101]	mRMR	30/343	71.0349%
DPC [28]	ANOVA	30/400	75.5569%
DDE [28]	ANOVA	30/400	67.0483%
TPC [26]	ANOVA	30/8000	75.5569%
PseAAC [33]	ANOVA	30/50	73.5075%
Geary [30]	mRMR	30/240	75.8706%
CKSAAP ($k = 0$~5)	ANOVA	30/2400	75.9282%
CKSAAP ($k = 0$)	ANOVA	30/400	75.7776%
CKSAAP ($k = 1$)	ANOVA	30/400	76.0885%
CKSAAP ($k = 2$)	ANOVA	30/400	75.7147%
CKSAAP ($k = 3$)	ANOVA	30/400	76.0878%
CKSAAP ($k = 4$)	ANOVA	30/400	75.8708%
CKSAAP ($k = 5$)	ANOVA	30/400	75.8701%

During the course of algorithm implementation, the open-source package libSVM supplied by Chang and Lin was used to implement the SVM algorithm [64]. Two parameters, c and γ, related to loss function and kernel function, respectively, were optimized by the method of gridding search using 6-fold cross-validation.

2.5. Performance Evaluation. Overfitting is an inevitable problem in machine learning. To reduce the influence of overfitting on model training, jackknife cross-validation or n-fold cross-validation is used to examine the power of the model on the training set [65]. The jackknife cross-validation method divides the training set into k subsets randomly, one of which is used to verify the accuracy of the model, and the other k-1 subsets are used to train the model. This method can avoid overfitting by generalizing the model with k-times repetition and is widely used in the machine learning process of small sample size data.

The performance of each model can be measured in terms of accuracy (ACC), sensitivity (SE), and specificity (SP) [66–72]. A confusion matrix can be set up with the help of the classification results, which further classifies the classification results of a binary classifier into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [73, 74]. These metrics are usually adopted to evaluate prediction quality [75–89]. Based on this, the parameters above can be expressed as follows:

$$\begin{cases} \mathrm{ACC} = \dfrac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} * 100\%, \\ \mathrm{SP} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} * 100\%, \\ \mathrm{SE} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} * 100\%, \end{cases} \tag{4}$$

where ACC is used to evaluate the overall performance of the model and SE and SP are used to measure the predictive ability of the model for positive and negative cases. Higher values of these parameters represent a better prediction performance of the model.

In addition, the receiver operating characteristic (ROC) curve is applied to evaluate the performance of the model further [90–100]. ROC curves are used to illustrate the diagnostic ability of a binary classifier, which shows the changes of SP and SE with varied thresholds. The area under the ROC curve (AUC) can be used to determine which classifier performs better in a quantitative way. ROC curve analysis can reflect the real performance of the model, especially for an unbalanced dataset.

3. Results and Discussion

3.1. Comparison of Feature Extraction Methods. We first compared the performance of common feature extraction methods on the training set identified by the SVM classifier. Feature vectors with high dimensions were selected by ANOVA or mRMR methods, depending on which method could maximize accuracy. The features of the sequences were extracted by the iFeature toolkit [44] and were then selected and classified using MATLAB and libSVM. The accuracies of the various methods are shown in Supplementary Materials (available here), calculated by 6-fold cross-validation. We found that AAC and composition, transition, and distribution (CTD) descriptors can classify human enzymes accurately, with an accuracy from 74.4% to 75.9%, and that AAC can achieve the highest accuracy, which means the frequency of all 20 amino acids can provide the most useful information about human enzyme classification, and thus, more useful information can be added to AAC to improve the model's prediction performance.

Based on the above discussion, other descriptors can be added to AAC to improve the model. The results of the predicted accuracy using different added descriptors are shown in Table 1, where the feature selection technique in ANOVA and mRMR with higher accuracy was used. The control variable method is used to find the optimal feature extraction method. Specifically, the dimension used for feature selection

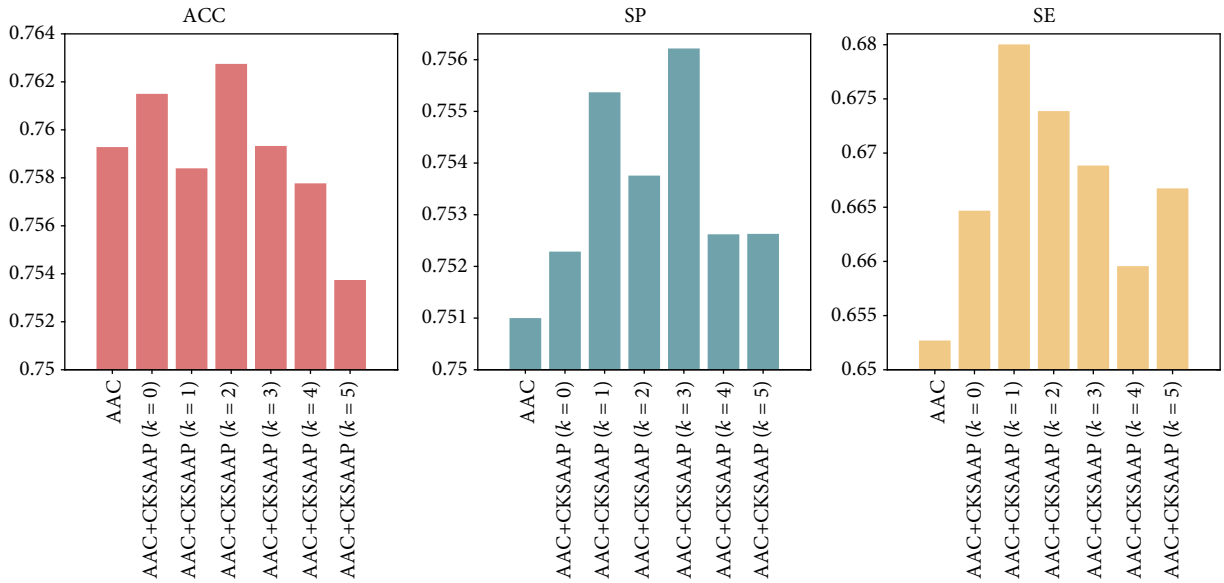

Figure 2: Comparison of SVM models trained by AAC alone versus AAC plus 6 types of CKSAAP.

is unchanged (30-D), and the performance of the SVM classifier under different feature extraction methods is compared to find the best feature extraction method for the identification of human enzymes. Based on the performance of the different descriptors on the training set, CKSAAP, which included not only information about the composition and sequence order but also information about the residue correlation, was determined to be the descriptor that can provide new valid information on the basis of AAC to improve the model performance.

3.2. Necessity of Feature Selection. Then, the performance of our method, using the AAC and CKSAAP descriptors as features, was measured in different dimensions that were selected to determine whether the feature selection method should be used to reduce redundant information and further improve the performance of our model. We employed AAC alone and AAC and 6 types of CKSAAP together as the predictor to train the SVM model. The results are presented in Figure 2. Relative to SE, SP, the ACC model using all of the features of AAC and CKSAAP was not much improved compared to using AAC alone and was even decreased, in spite of features in CKSAAP that include useless information that influences the precision of our model. This result could lead to the conclusion that a feature selection technique is necessary to reduce redundant information and improve the precision of our model.

3.3. Selection of Significant Features. After determining the feature selection techniques necessary to improve the prediction accuracy of the model, the size of the significant features of the CKSAAP descriptors that we selected needed to be identified. We used ANOVA to select informative k-spaced amino acid pairs. The definite means are as follows: (1) Evaluate all of the amino acid pairs and sort them according to the difference between the two types of amino acids. (2) Each CKSAAP feature is sequentially added to the parameter subset with AAC according to the sorted order. (3) The SVM-based model is trained using the parameter subset. Then, all of the results are compared to find the best feature subset of the significant features we selected.

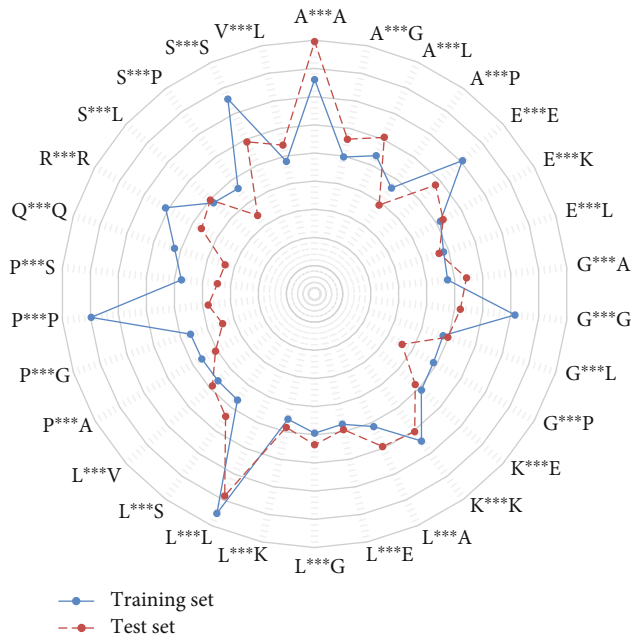

Figure 3: Results of the top 30 feature parameters of CKSAAP ($k = 3$). The radius of each point indicates the variance of the feature parameter in the training set or test set.

According to these methods, taking $k = 3$ as an example, the top 30 feature parameters of CKSAAP were selected and are shown as Figure 3, and the variance of 50 feature parameters in both the training and test sets are also shown. A***A and L***L have a large variance in both the training and test sets, foreshadowing that they contain more information.

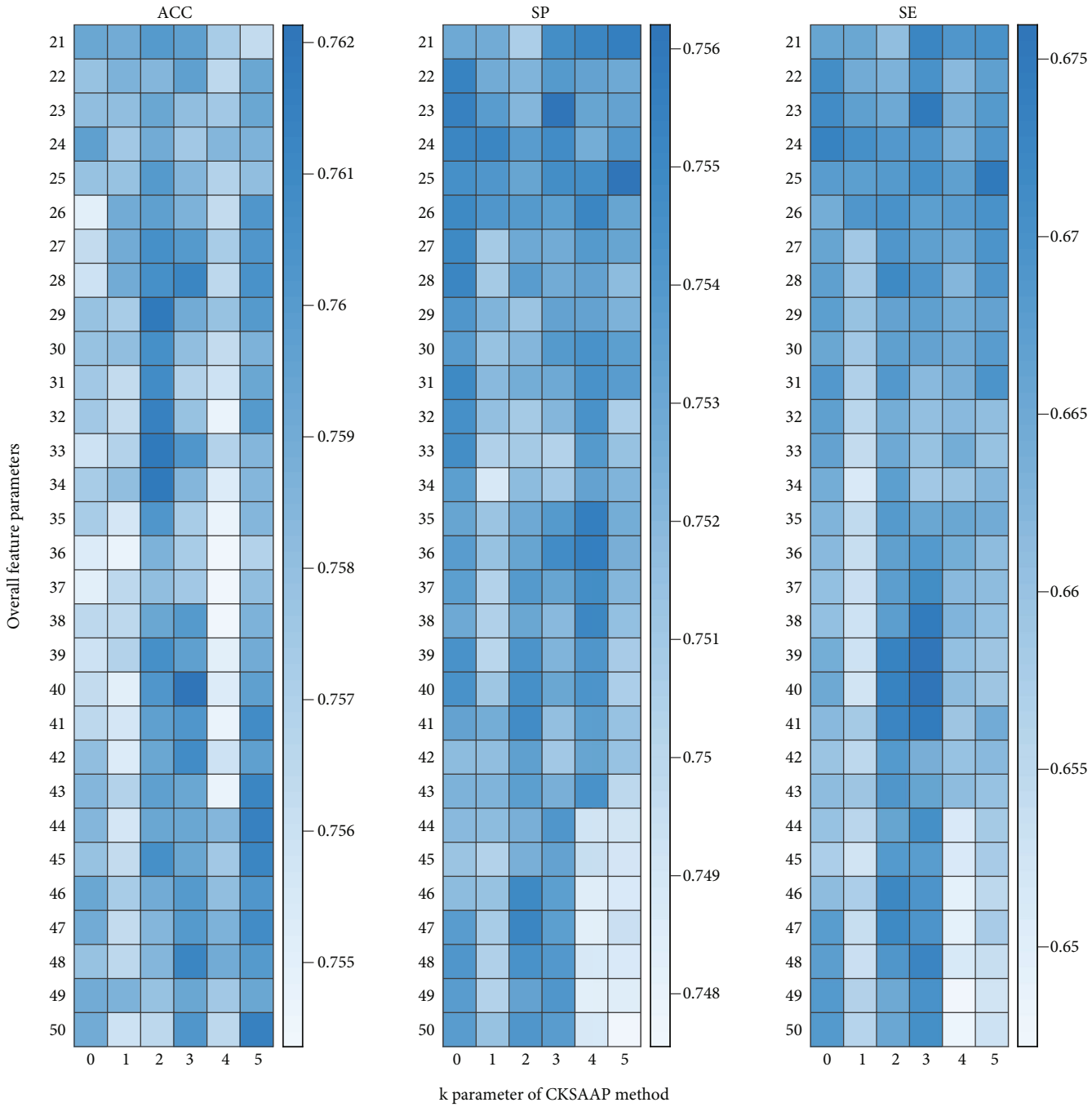

Figure 4: Results of ACC, SP, and SE of the model trained by 20 AAC parameters and 1–30 important CKSAAP parameters selected by the ANOVA technique.

We used the top 30 feature parameters of CKSAAP from ANOVA added into the AAC parameters to train the model, change the value of k during feature extraction, and change the number of features added to AAC at the same time to select the model with the best performance, instead of only changing the feature extraction method, and the results are shown in Figure 4. We obtained a maximum accuracy when we used 20 AAC parameters and 20 CKSAAP parameters ($k = 3$) for 40 feature parameters overall. The c/γ values used in the SVM-based model are 1.1487 ($2^{0.2}$) and 147.0334 ($2^{7.2}$), respectively. The accuracy reached 76.2135%, and SP and SE reached 0.7530 and 0.6760, respectively, which are all higher than the accuracy achieved in past research. We also measured the performance of the above model by making predictions on the test set and obtained an overall accuracy of 76.4585%, which indicates that the SVM model we established performs well in the classification of human enzymes. The 20 informative 3-spaced amino acid pairs that are used in the model training stage are L***L, P***P, A***A, S***S, G***G, E***E, K***K, R***R, A***L, Q***Q, E***K, L***A, K***E, A***G, L***G, G***P, S***L,

Table 2: Comparison of the performance of various mainstream classifiers and the classifier implemented in our paper. ACC, SP, and SE of different classifiers on both the training set and the test set are compared.

Classifiers	Training set			Test set		
	ACC	SP	SE	ACC	SP	SE
This work (SVM)	76.2135%	0.753	0.676	76.4585%	0.762	0.657
Naive Bayes	61.0697%	0.466	0.833	65.7625%	0.507	0.794
Random Forest	74.3781%	0.703	0.454	74.7691%	0.710	0.472
Logistic	69.5274%	0.598	0.374	68.4237%	0.587	0.329
KNN	62.8420%	0.474	0.646	63.0502%	0.480	0.658
Ensembles for Boosting	69.6206%	0.588	0.420	68.6796%	0.573	0.411

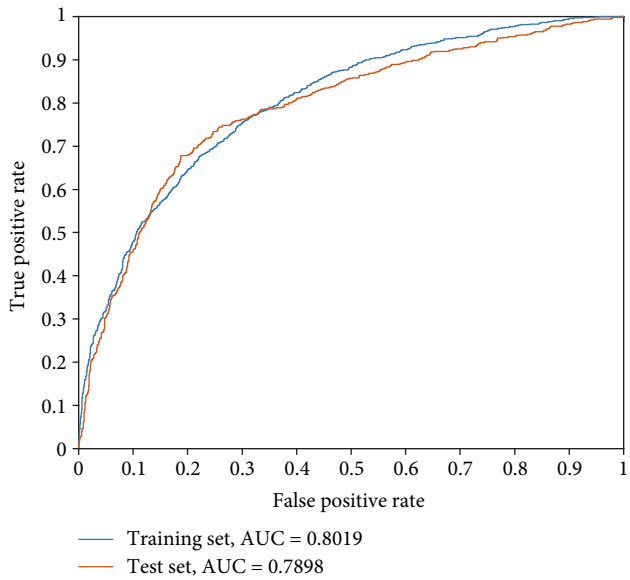

Figure 5: The ROC curves of our model on both the training set and test set, with AUCs of 0.8019 and 0.7898, respectively.

E***L, V***L, and G***L (* indicates the other characters between two amino acids, i.e., the space), which may play important roles in human enzymes.

Furthermore, various mainstream classifiers, i.e., Naive Bayes, Random Forest, Logistic, *K*-nearest neighbor (KNN), and Ensembles for Boosting [102–105] are compared with our model in both the training set and the test set using 6-fold cross-validation in Table 2, and the result shows that the SVM-based classifier in our paper performs best. In addition, the ROC curve of our model performed well on both the training set and the test set, as shown in Figure 5, which confirms the classification effect of the model. The AUC reached 0.8019 and 0.7898 in the training set and the test set, respectively, demonstrating that our method for human-enzyme classification is effective and that more accurate classification results can now be obtained.

4. Conclusion

In this study, we proposed an effective and novel method to identify human enzymes using AAC and CKSAAP that is based on short-range interactions of amino acid pairs rather than the physicochemical properties of the sequences. By using ANOVA to select informative feature parameters, 20 amino acid pairs in 3-space are selected to add 20 residues and feed their frequency into an SVM classifier. The jackknife cross-validated accuracy was 76.46% in the training set, demonstrating that fewer feature parameters were used and a higher accuracy was reached compared to previous research. Moreover, we compared the performance of the model using different feature extraction methods, and the results showed that residue-frequency-based methods perform better than other methods, and a web server based on our method will be implemented in the future. In addition, some important feature parameters selected by ANOVA, e.g., A***A and L***L, may contain vital information in regard to the identification of human enzymes, which we hope to discuss more deeply in the future.

Data Availability

In our experiment, the sequence data of the training set and the feature vectors of both the training set and the test set extracted by the iFeature toolkit are available online at https://github.com/Fu-Zhang/Identification-of-human-enzymes. The sequence data of the test set are available in the Supplementary Materials of Reference [11].

Conflicts of Interest

The authors have declared no competing interests.

Authors' Contributions

YZ and LJ conceived and designed the project. LZ and BD conducted experiments and analyzed the data. LZ and LJ wrote the paper. ZT and YZ revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No. 61901103) and the Natural Science Foundation of Heilongjiang Province (No. LH2019F002).

Supplementary Materials

Accuracy of model training with various feature extraction methods by 6-fold cross-validation. Two feature selection methods, ANOVA and mRMR, are used and the feature selection method with higher accuracy is selected and included in the table. *(Supplementary Materials)*

References

[1] *Nomenclature E, Webb E*, Academic Press, San Diego, CA, USA, 1992.

[2] J.-X. Tan, H. Lv, F. Wang, F.-Y. Dao, W. Chen, and H. Ding, "A survey for predicting enzyme family classes using machine learning methods," *Current Drug Targets*, vol. 20, no. 5, pp. 540–550, 2019.

[3] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, p. 158, 2018.

[4] B. Liu, Y. Zhu, and K. Yan, "Fold-LTR-TCP: protein fold recognition based on triadic closure principle," *Briefings in Bioinformatics*, 2019.

[5] L. Wei, Q. Zou, M. Liao, H. Lu, and Y. Zhao, "A novel machine learning method for cytokine-receptor interaction prediction," *Combinatorial Chemistry & High Throughput Screening*, vol. 19, no. 2, pp. 144–152, 2016.

[6] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395–2402, 2019.

[7] K. Qu, L. Wei, and Q. Zou, "A review of DNA-binding proteins prediction methods," *Current Bioinformatics*, vol. 14, no. 3, pp. 246–254, 2019.

[8] L. J. Jensen, M. Skovgaard, and S. Brunak, "Prediction of novel archaeal enzymes from sequence-derived features," *Protein Science*, vol. 11, no. 12, pp. 2894–2898, 2002.

[9] K.-C. Chou and Y.-D. Cai, "Using GO-PseAA predictor to predict enzyme sub-class," *Biochemical and Biophysical Research Communications*, vol. 325, no. 2, pp. 506–509, 2004.

[10] C. Z. Cai, L. Y. Han, Z. L. Ji, and Y. Z. Chen, "Enzyme family classification by support vector machines," *Proteins: Structure, Function, and Bioinformatics*, vol. 55, no. 1, pp. 66–76, 2004.

[11] Y.-D. Cai and K.-C. Chou, "Using functional domain composition to predict enzyme family classes," *Journal of Proteome Research*, vol. 4, no. 1, pp. 109–111, 2005.

[12] Y.-D. Cai and K.-C. Chou, "Predicting enzyme subclass by functional domain composition and pseudo amino acid composition," *Journal of Proteome Research*, vol. 4, no. 3, pp. 967–971, 2005.

[13] H.-B. Shen and K.-C. Chou, "EzyPred: a top-down approach for predicting enzyme functional classes and subclasses," *Biochemical and Biophysical Research Communications*, vol. 364, no. 1, pp. 53–59, 2007.

[14] E. Nasibov and C. Kandemir-Cavas, "Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction," *Computational Biology and Chemistry*, vol. 33, no. 6, pp. 461–464, 2009.

[15] R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto et al., "Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of *Leishmania* proteins," *Journal of Proteome Research*, vol. 8, no. 9, pp. 4372–4382, 2009.

[16] J.-D. Qiu, J.-H. Huang, S.-P. Shi, and R.-P. Liang, "Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform," *Protein and Peptide Letters*, vol. 17, no. 6, pp. 715–722, 2010.

[17] R. Shi and X. Hu, "Predicting enzyme subclasses by using support vector machine with composite vectors," *Protein and Peptide Letters*, vol. 17, no. 5, pp. 599–604, 2010.

[18] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[19] B. Niu, Y. Lu, J. Lu et al., "Prediction of Enzyme's family based on protein-protein interaction network," *Current Bioinformatics*, vol. 10, no. 1, pp. 16–21, 2015.

[20] Y. Li, S. Wang, R. Umarov et al., "DEEPre: sequence-based enzyme EC number prediction by deep learning," *Bioinformatics*, vol. 34, no. 5, pp. 760–769, 2018.

[21] S. A. Memon, K. A. Khan, and H. Naveed, "Enzyme function prediction using deep learning," *Biophysical Journal*, vol. 118, no. 3, article 533a, 2020.

[22] Y. Wu, H. Tang, W. Chen, and H. Lin, "Predicting human enzyme family classes by using pseudo amino acid composition," *Current Proteomics*, vol. 13, no. 2, pp. 99–104, 2016.

[23] T. U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.

[24] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[25] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2018.

[26] M. Bhasin and G. P. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, 2004.

[27] B. Liu, "BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1280–1294, 2019.

[28] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor," *OMICS: A Journal of Integrative Biology*, vol. 19, no. 10, pp. 648–658, 2015.

[29] H. Tang, Y. W. Zhao, P. Zou et al., "HBPred: a tool to identify growth hormone-binding proteins," *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 957–964, 2018.

[30] R. R. Sokal and B. A. Thomson, "Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population," *American Journal of Physical Anthropology*, vol. 129, no. 1, pp. 121–131, 2006.

[31] C. Cai, L. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3692–3697, 2003.

[32] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy*

of Sciences of the United States of America, vol. 92, no. 19, pp. 8700–8704, 1995.

[33] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[34] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[35] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 462, pp. 230–239, 2019.

[36] B. Liu, X. Gao, and H. Zhang, "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 Research*, vol. 47, no. 20, article e127, 2019.

[37] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

[38] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowledge-Based Systems*, vol. 163, pp. 787–793, 2019.

[39] B. Yu, W. Qiu, C. Chen et al., "SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting," *Bioinformatics*, vol. 36, 2019.

[40] X. Zhao, Q. Jiao, H. Li et al., "ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles," *BMC Bioinformatics*, vol. 21, no. 1, p. 43, 2020.

[41] K. Chen, Y. Jiang, L. Du, and L. Kurgan, "Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs," *Journal of Computational Chemistry*, vol. 30, no. 1, pp. 163–172, 2009.

[42] K. Chen, L. Kurgan, and M. Rahbari, "Prediction of protein crystallization using collocation of amino acid pairs," *Biochemical and Biophysical Research Communications*, vol. 355, no. 3, pp. 764–769, 2007.

[43] Z. Chen, Y. Chen, X. Wang, C. Wang, R. Yan, and Z. Zhang, "Prediction of ubiquitination sites by using the composition of *k*-spaced amino acid pairs," *PLoS One*, vol. 6, no. 7, article e22930, 2011.

[44] Z. Chen, P. Zhao, F. Li et al., "iFeature: a python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.

[45] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[46] S. P. Wang, Q. Zhang, J. Lu, and Y. D. Cai, "Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm," *Current Bioinformatics*, vol. 13, no. 1, pp. 3–13, 2018.

[47] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[48] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015.

[49] W. Yang, X.-J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-Golgi localization," *Current Bioinformatics*, vol. 14, pp. 234–240, 2019.

[50] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 283–291, 2019.

[51] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 687–695, 2017.

[52] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism," *Bioinformatics*, vol. 36, 2019.

[53] J. X. Tan, S. H. Li, Z. M. Zhang et al., "Identification of hormone binding proteins based on machine learning methods," *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2466–2480, 2019.

[54] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso," *Journal of Theoretical Biology*, vol. 486, 2020.

[55] Y. Wang, F. Shi, L. Cao et al., "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinformatics*, vol. 14, no. 4, pp. 282–294, 2019.

[56] X. Du, X. Li, W. Li, Y. Yan, and Y. Zhang, "Identification and analysis of cancer diagnosis using probabilistic classification vector machines with feature selection," *Current Bioinformatics*, vol. 13, no. 6, pp. 625–632, 2018.

[57] N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, "Discriminating Ramos and Jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine," *Current Bioinformatics*, vol. 13, pp. 50–56, 2018.

[58] Q. H. Jiang, G. H. Wang, S. L. Jin, Y. Li, and Y. D. Wang, "Predicting human microRNA-disease associations based on support vector machine," *International Journal of Data Mining and Bioinformatics*, vol. 8, no. 3, pp. 282–293, 2013.

[59] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, p. 3140, 2018.

[60] B. Liu, C. Li, and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in Bioinformatics*, 2019.

[61] Y. Qiao, Y. Xiong, H. Gao, X. Zhu, and P. Chen, "Protein-protein interface hot spots prediction based on a hybrid feature selection strategy," *BMC Bioinformatics*, vol. 19, no. 1, p. 14, 2018.

[62] L. Cheng, H. Yang, H. Zhao et al., "MetSigDis: a manually curated resource for the metabolic signatures of diseases," *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 203–209, 2019.

[63] L. Cheng, H. Zhuang, S. Yang, H. Jiang, S. Wang, and J. Zhang, "Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: a Mendelian randomization study," *Frontiers in Genetics*, vol. 9, p. 657, 2018.

[64] C.-C. Chang and C.-J. Lin, "LIBSVM," *Acm Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[65] H. Ding, W. Yang, H. Tang et al., "PHYPred: a tool for identifying bacteriophage enzymes and hydrolases," *Virologica Sinica*, vol. 31, no. 4, pp. 350–352, 2016.

[66] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "deepDR: a network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[67] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "A polar-metric-based evolutionary algorithm," *IEEE Transactions on Cybernetics*, pp. 1–12, 2020.

[68] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods," *Briefings in Bioinformatics*, 2019.

[69] X. Zeng, S. Zhu, W. Lu et al., "Target identification among known drugs by deep learning from heterogeneous networks," *Chemical Science*, 2020.

[70] C. Chen, Q. Zhang, Q. Ma, and B. Yu, "LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 54–64, 2019.

[71] G. Wang, Y. Wang, W. Feng et al., "Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells," *BMC Genomics*, vol. 9, Supplement 2, p. S22, 2008.

[72] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells," *PLoS One*, vol. 5, no. 7, article e11794, 2010.

[73] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Transactions on Cybernetics*, pp. 1–12, 2019.

[74] X. Liu, Z. Hong, J. Liu et al., "Computational methods for identifying the critical nodes in biological networks," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 486–497, 2020.

[75] C. Shen, Y. Ding, J. Tang, L. Jiang, and F. Guo, "LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information," *IEEE Access*, vol. 7, pp. 13486–13496, 2019.

[76] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via semi-supervised model and multiple kernel learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2619–2632, 2019.

[77] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[78] L. Wei, S. Wan, J. Guo, and K. K. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artificial Intelligence in Medicine*, vol. 83, pp. 82–90, 2017.

[79] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[80] B. Liu and K. Li, "iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 80–87, 2019.

[81] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*," *Briefings in Functional Genomics*, vol. 18, no. 6, pp. 367–376, 2019.

[82] X. Shan, X. Wang, C. D. Li et al., "Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method," *Journal of Chemical Information and Modeling*, vol. 59, no. 11, pp. 4577–4586, 2019.

[83] Y. Chu, A. C. Kaushik, X. Wang et al., "DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Briefings in Bioinformatics*, 2019.

[84] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: pseudouridine sites identification based on RNA sequence information," *BMC Bioinformatics*, vol. 19, no. 1, p. 306, 2018.

[85] L. Cheng, H. Zhao, P. Wang et al., "Computational methods for identifying similar diseases," *Molecular Therapy - Nucleic Acids*, vol. 18, pp. 590–604, 2019.

[86] L. Cheng, "Computational and biological methods for gene therapy," *Current Gene Therapy*, vol. 19, no. 4, pp. 210–210, 2019.

[87] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in *Homo sapiens*," *Briefings in Bioinformatics*, 2020.

[88] G. Wang, X. Luo, J. Wang et al., "MeDReaders: a database for transcription factors that bind to methylated DNA," *Nucleic Acids Research*, vol. 46, no. D1, pp. D146–D151, 2018.

[89] L. Cheng, P. Wang, R. Tian et al., "LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Research*, vol. 47, no. D1, pp. D140–D144, 2019.

[90] F. Hsieh and B. W. Turnbull, "Nonparametric and semiparametric estimation of the receiver operating characteristic curve," *Annals of Statistics*, vol. 24, no. 1, pp. 25–40, 1996.

[91] L. Wei, J. Hu, F. Li, J. Song, R. Su, and Q. Zou, "Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms," *Briefings in Bioinformatics*, vol. 18, 2018.

[92] B. Liu and Y. Zhu, "ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into Learning to Rank," *IEEE ACCESS*, vol. 7, pp. 102499–102507, 2019.

[93] T. Fang, Z. Zhang, R. Sun et al., "RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition," *Molecular Therapy - Nucleic Acids*, vol. 18, pp. 739–747, 2019.

[94] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D. Q. Wei, "PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers in Microbiology*, vol. 9, p. 2571, 2018.

[95] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.

[96] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2020.

[97] H. Y. Lai, Z. Y. Zhang, Z. D. Su et al., "iProEP: a computational predictor for predicting promoter," *Molecular Therapy-Nucleic Acids*, vol. 17, pp. 337–346, 2019.

[98] X. Sun, T. Jin, C. Chen, X. Cui, Q. Ma, and B. Yu, "RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net," *Chemometrics and Intelligent Laboratory Systems*, vol. 197, article 103919, 2020.

[99] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification in *Arabidopsis* using multiple histone markers," *BioMed Research International*, vol. 2015, Article ID 861402, 10 pages, 2015.

[100] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of microRNA promoter prediction and transcription factor mediated regulatory network," *BioMed Research International*, vol. 2017, Article ID 7049406, 8 pages, 2017.

[101] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[102] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 215, 2019.

[103] Q. Ning, Z. Ma, and X. Zhao, "dForml(KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components," *Journal of Theoretical Biology*, vol. 470, pp. 43–49, 2019.

[104] H. Hu, L. Zhang, H. Ai et al., "HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy," *RNA Biology*, vol. 15, no. 6, pp. 797–806, 2018.

[105] S. Tagore, A. Gorohovski, L. J. Jensen, and M. Frenkel-Morgenstern, "ProtFus: a comprehensive method characterizing protein-protein interactions of fusion proteins," *PLoS Computational Biology*, vol. 15, no. 8, article e1007239, 2019.