Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 8926750, 9 pages
https://doi.org/10.1155/2020/8926750

Research Article

A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD

Zhiyu Tao, Yanjuan Li, Zhixia Teng, and Yuming Zhao

Information and Computer Engineering College, Northeast Forestry University, Harbin 150040, China

Correspondence should be addressed to Yuming Zhao; zym@nefu.edu.cn

Zhiyu Tao and Yanjuan Li contributed equally to this work.

Received 4 August 2020; Revised 14 August 2020; Accepted 16 September 2020; Published 19 October 2020

Academic Editor: Hui Ding

Copyright © 2020 Zhiyu Tao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of computer technology, many machine learning algorithms have been applied to the field of biology, forming the discipline of bioinformatics. Protein function prediction is a classic research topic in this subject area. Though many scholars have made achievements in identifying protein by different algorithms, they often extract a large number of feature types and use very complex classification methods to obtain little improvement in the classification effect, and this process is very time-consuming. In this research, we attempt to utilize as few features as possible to classify vesicular transportation proteins and to simultaneously obtain a comparative satisfactory classification result. We adopt CTDC which is a submethod of the method of composition, transition, and distribution (CTD) to extract only 39 features from each sequence, and LibSVM is used as the classification method. We use the SMOTE method to deal with the problem of dataset imbalance. There are 11619 protein sequences in our dataset. We selected 4428 sequences to train our classification model and selected other 1832 sequences from our dataset to test the classification effect and finally achieved an accuracy of 71.77%. After dimension reduction by MRMD, the accuracy is 72.16%.

1. Introduction

Protein, regarded as the material basis of life and the caretaker of life activities [1], participates in all the functions of maintaining individual survival, including catalyzing specific biochemical reactions and participating in immune response. The protein diversity is increased by alternative splicing and posttranslation modifications [2, 3]. Hence, the topic of protein function prediction came into being around the time of the birth of bioinformatics [4–11]. In view of the different functions of protein, there are various kinds to be classified [12–17]. Many scholars are devoted to the classification of different functions of an enzyme [18–23], and some apply themselves to the recognition of whether a protein sequence is an effecter protein. In this thesis, we attempt to determine if a protein is a vesicular transport protein.

Substances with small molecular weight, such as water or ions, will directly pass through the cell membrane by free diffusion or through the ion channels embedded in the cell membrane. However, macromolecular materials like proteins cannot directly pass through the cell membrane. In the process of transportation in and out of the cell, they are first surrounded by a layer of membrane generated by cell-forming vesicles and then through the fusion or rupture of vesicles with the cell membrane or various organelle membranes to complete material transportation. This process is called vesicular transport. The key role to facilitate this process is vesicular transporter, which is a kind of ubiquitous protein in the cell membrane and organelle membrane. When macromolecular materials are to be transported across the membrane, a specific vesicle transport will concentrate them or supervise the specific organelles to produce different vesicle structures to carry or to wrap the materials to be transmitted. Vesicle transport activity occurs widely between cells or within cells, such as the transmission of neurotransmitters between nerve cells and the operation of the immune system, which is essential for maintaining life. In the field of biology, there have been many advanced studies on cell

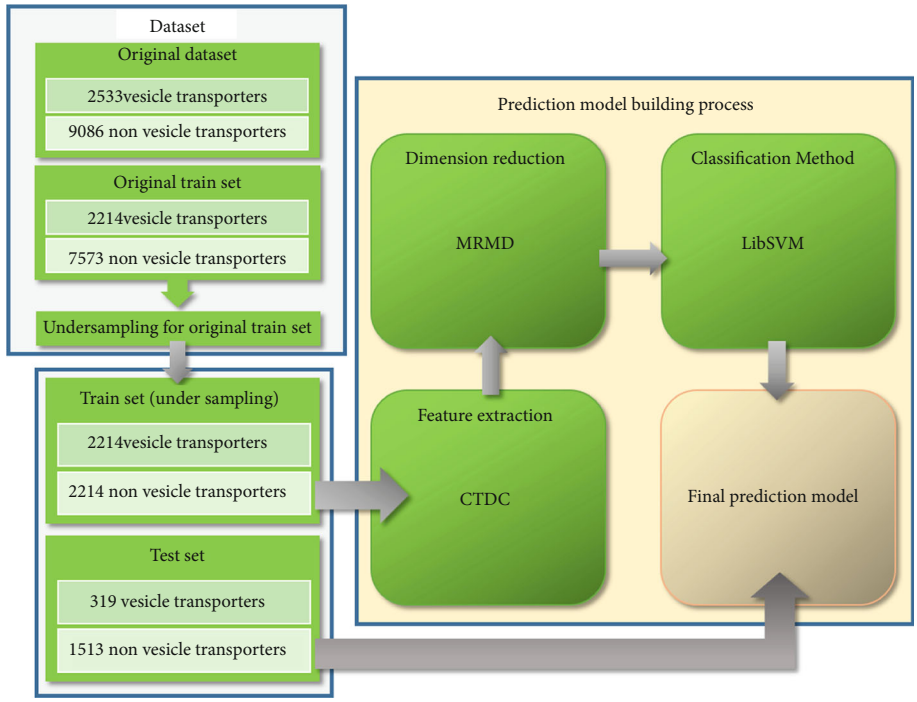

Figure 1: Flowchart of identifying vesicular transporters.

vesicle transport, and the research areas are also diverse. For example, Rothman et al. [24] studied the problem about the transport of proteins in Golgi matrix, the composition, and structure of Golgi-coated vesicles. Liu et al. [25] concentrated on research about the effect vesicular transporter that plays in synaptic transmission and neurodegeneration. Similarly, many human diseases are also related to the abnormal action of vesicular transport in cells. Brain dopamine-serotonin vesicular transport disease, which can cause movement disorder in infancy, is closely related to vascular monoamine transporter 2 (VMAT2) [26]. In addition, many similar examples are constantly discovered. Increasingly, more diseases are associated with gene mutations, which are responsible for the vesicular transport function.

With the development of this field, an increasing number of vesicular transport proteins and other proteins have been found. There is growing desire for rapid identification of vesicular transporters, which is difficult to meet with biological technology. This type of research requires bioinformatics scholars to use machine learning and other computational methods to process and to analyze massive protein sequences. Thus far, the research on using computational methods to identify vesicle transporters is scant. In 2019, Le et al. [27] used PSSM matrix to store sequence features and convolutional neural network (CNN) to determine whether the sequence is a SNARE protein, which is a kind of vesicular transporter. In the same year, these authors used a classifier called GRU based on CNN to identify vesicular transporters. However, for the identification of protein, DNA and RNA, the process to deal with these problems is similar. In recent years, the two steps of the process, feature extraction and classification, have become increasingly complex, and this is also true in the field of identifying vesicular transporters. Meanwhile, we try as much as possible to use a simpler way of feature extraction and classification, to ensure a better classification effect. Finally, we use the composition descriptor in the composition, transition, and distribution (CTDC) and LibSVM as the methods of feature extraction and classification, which are widely used by many scholars. It is a novel idea about our research that the feature dimension of our final prediction process is reduced to 29. Our flowchart is shown in Figure 1.

2. Materials and Methods

2.1. Dataset. Our data come from the previous research of Le et al. These data have been processed by BLAST to ensure that the similarity between any two sequences is less than 30%. In addition, we use random undersampling in order to balance the number of positive and negative samples in the training set. Finally, there are 4428 sequences in the final training set and 1832 sequences in the test set. Table 1 details the composition of the dataset.

2.2. Method to Feature Extraction. Feature extraction is very important for constructing a predictor [28–37]. We use the CTDC method in iLearn toolkit to extract features of protein sequences. Developed from iFeatures, iLearn is a comprehensive toolkit based on Python, which was designed by Chen et al. that can be downloaded at http://ilearn.erc.monash.edu. As a powerful platform, it not only integrates a series of feature extraction and analysis methods but provides many machine learning algorithms for classification. CTDC is the first part of the CTD feature method in iLearn based on the first of three descriptors.

Table 1: The dataset used in this study.

	Original	Original train set	Train set	Test set
Vesicular transport	2533	2214	2214	319
Nonvesicular transport	9086	7573	2214	1513

CTD is a classic sequence feature extraction method that was first proposed by Dubchak et al. [38] in 1995. It consists of three descriptors: composition (C), transition (T), and distribution (D). Composition refers to the ratio of the number of single amino acids with specific properties (or several small amino acid sequence fragments with certain physical and chemical properties) in the whole sequence [39]. Composition can be expressed with the following formula:

$$\text{Composition} = \frac{n_x}{N}(x = a, b, c \cdots), \tag{1}$$

where x represents amino acids with specific groups or sequence fragments with special physical and chemical properties, and a, b, and c represent different kinds of groups. N represents the total length of the sequence. The second descriptor represents the ratio of two closely adjacent groups to the total sequence calculated as

$$\text{transition} = \frac{yx + xy}{N - 1}(x = a, b, c \cdots y = a, b, c \cdots). \tag{2}$$

In Eq. (2), xy and yx denote two closely adjacent groups. The third descriptor, distribution, represents the general spreading state of special groups in the whole sequence. From the first amino acid of the sequence, calculate the proportion of an amino acid carrying a specific group in five subchains for all the amino acids in a sequence. These five chains contain the first, 25%, 50%, 75%, and 100% special amino acid from the first amino acid of the sequence.

After our experiment, the features extracted from transition and distribution contributed little to the classification effect, so we only select the features extracted from composition. In iLearn, there are 13 kinds of physicochemical properties adopted, and each physicochemical property has three kinds of amino acid combination patterns. The concrete meaning of these properties comes from the research results of Tomii and Kanehisa [40] in 1996.

2.3. Method for Classification. We use LibSVM method based on Weka. LibSVM is a library about the support vector machine (SVM) developed by Professor Lin et al. in 2001. It has been widely used in bioinformatics [41–54]. It has the advantages of being a small program that is flexible, with less inputting parameters, is open source to expand easily, and thus has become the most widely used SVM Library in China. This library tool can be accessed at https://www.csie.ntu.edu.tw/~cjlin/. Weka is a free and noncommercial mining platform, which has a series of functional modules that basically meet various needs in data analysis, such as a variety of different classification and regression algorithms and performing cross validation during classification, automatically. LibSVM classification has been supported since Weka version 3.5.

SVM is a kind of generalized linear classifier that relies on supervised learning [55–60]. The key to classification is to form a hyperplane in multidimensional feature space through algorithm calculation, which can approximate separate positive and negative samples; it can be expressed mathematically as

$$\omega^T X_i + b = 0, \tag{3}$$

In (3), X is a vector composed of coordinate values of any point on the hyperplane in each dimension and ω is a vector that we need to calculate. In addition, in order to make the sum of the distance between the positive and negative sample set and the hyperplane reach the farthest, we need to construct two planes parallel to the hyperplane as the interval boundary to distinguish the sample classification. However, in most cases, positive and negative samples cannot be completely divided on both sides of a plane, so generally we will allow some samples to be divided incorrectly, which we called soft interval. Finally, the problem is simplified to formula (4).

$$\min_{\omega,b,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i s.t. y_i\left(\omega^T x_i + b\right) \geq 1 - \xi_i, i = 1, 2, \cdots, m, \xi_i \geq 0, \tag{4}$$

where ξ_i represents a relaxation variable for each sample point, and C is the penalty parameter that needs to be set manually according to the actual situation. The Lagrange function corresponding to formula (5) can be shown as

$$L(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1 - \xi_i - y_i\left(\omega^T x_i + b\right)\right] + \sum_{i=1}^{m}\beta_i(-\xi_i), \tag{5}$$

where parameters α_i and β_i are Lagrange multipliers. At the same time, to solve the problem conveniently, we need to use the technique about Lagrangian duality and set the kernel function. The dual problem to Lagrange function is represented as

$$\max_{\alpha,\beta}\min_{\omega,b,\xi} L(\omega, b, \xi, \alpha, \beta)\, s.t. \alpha_i \geq 0, \beta_i \geq 0\ i = 1, 2, \cdots, m. \tag{6}$$

In this experiment, the radial basis function (RBF) is adopted as the kernel function, which is also the default setting in LibSVM. Two parameters, the cost (c) and gamma (g), need to be determined before building the classification model by using Weka. The parameter c is called the penalty coefficient. The higher the value of c is, the easier it is to over fit. And g is a parameter of RBF function after it is selected as kernel which affects the speed of process of training and

prediction. There is no universally recognized best method for parameter selection, and the common method is to let c and g take values within a certain range, and then set different c and g in the process of training set data classification. Finally, use cross validation to get the classification accuracy verified by the training set in this groups c and g, and select the group with the best classification result by comparison [61]. It is a complicated process, but in LibSVM toolkit, the parameter optimization is automated, and it no longer needs to be manually adjusted. We use the program, grid.py in the LibSVM tool folder to get the optimal parameters.

2.4. MRMD for Dimensionality Reduction. The max-relevance-max-distance (MRMD) is a dimensionality reduction algorithm, which was developed by Zou et al. [62, 63] in 2015 that can be downloaded at https://github.com/heshida01/mrmd/tree/master/mrmdjar. It is based on a series of distance functions to judge the feature independence. The process of dimensionality reduction consists of three steps. First, the contribution of each feature to classification is evaluated and then the contribution is quantified. Second, sort the features according to their contribution to the classification. Third, select different numbers of features in order to classify and then record the results. For example, select the first feature the first time, select the first two features the second time, etc., until the number of selected features reaches the maximum; that is, all features are selected, and the classification test stops. By comparing the results of these classification tests, the best group is selected, and the features selected in this group are retained and regarded as the result of dimension reduction.

MRMD algorithm analyzes the contribution of each feature to the prediction process mainly through two aspects, max relevance and max distance. Max relevance (MR) is used to calculate the Pearson correlation coefficient between features and samples to quantify the correlation between features and case classes. As shown in Formula (7), Pearson correlation coefficient is equal to the covariance divided by the product of their respective standard deviations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{7}$$

The vectors X and Y are composed of the ith feature from the sequence and the class label to which these sequences belong. Max distance (MD) is used to analyze the redundancy between features. Specifically, we calculate the three indexes between features:

$$\begin{aligned} \mathrm{ED}(X, Y) &= \sqrt{\sum_{k=1}^{N} (x_k - y_k)^2}, \\ \mathrm{COS}(X, Y) &= \frac{X \cdot Y}{\|X\| \cdot \|Y\|}, \\ \mathrm{TC}(X, Y) &= \frac{X \cdot Y}{\|X\|^2 + \|Y\|^2 - X \cdot Y}. \end{aligned} \tag{8}$$

In (8), these indexes are called Euclidean distance, Cosine similarity, and Tanimoto coefficient. The value of MD is obtained by comprehensive consideration of these three indexes.

Finally, the value of the contribution of the classification of each feature is obtained by adding the values of MR and MD in a certain proportion.

2.5. Evaluation of Classification Results. We adopt cross validation (CV) to evaluate the experimental results objectively. It is a classic, analytical method for judging the performance of a prediction model [64–78]. The core idea is to take out most of the samples in a given dataset to build a classification model, to leave a small part of the samples, to use the newly established model for prediction, and to calculate the forecast errors of these small samples and to record their sum of squares. This process continues until all samples are predicted once and only once. There are three common CV methods: hold-out method, K-fold cross validation (K-CV), and leave-one-out cross validation (LOO-CV). We take the second approach, K-CV.

In K-fold cross validation, the initial data are divided into k groups of subdatasets. A group of independent subdatasets are retained as the validation model data, and other k-1 subdatasets are used for training. In this way, we can get k models and take each prediction result of the classifier into account. In general, the operation is to take the average value of each index of every classification time from k models. The value of K can be set according to the actual situation, and here we set its value to 5. After 5-fold cross validation set in Weka, in order to evaluate the results of classification, some indexes are often used [79–85]. The metrics we use are recall, precision, MCC, and accuracy, and their corresponding formulas are as follows:

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP + FN}, \\ \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Accuracy} &= \frac{TP + TN}{TP + FN + FP + FN}, \\ \text{MCC} &= \frac{1 - ((FN/TP + FN) + (FP/TN + FP))}{\sqrt{(1 + (FP - FN/TP + FN))(1 + (FN - FP/TN + FP))}}. \end{aligned} \tag{9}$$

For the convenience of description, we use "positive" to represent vesicular transporters and "negative" to represent nonvesicular transporters. In (7), the letter T means true (correct). The letter N means false (incorrect). P is the positive sample, and N represents the negative sample. For example, TP means that the positive samples are correctly identified.

3. Results and Discussion

After optimizing the parameters of the dataset having the whole 39 dimensional features extracted by CTDC, we first

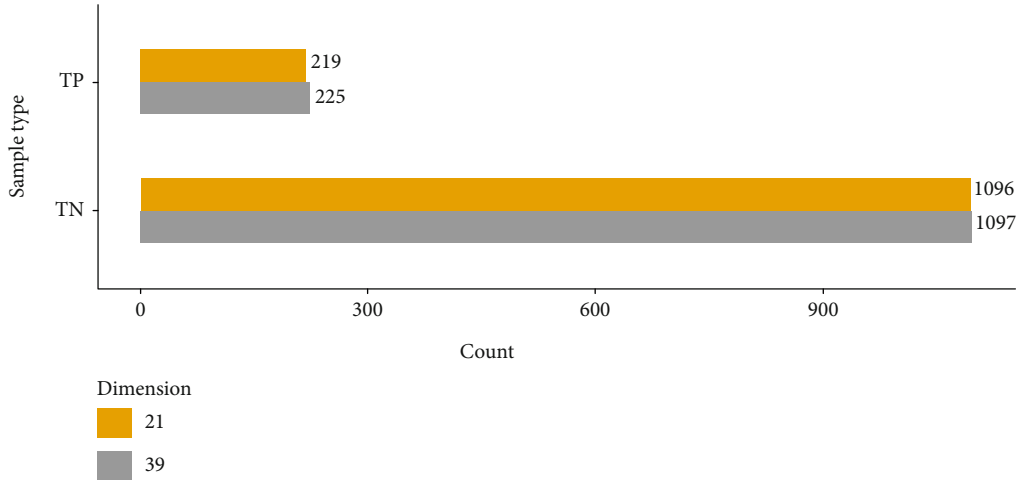

Figure 2: Number of samples correctly predicted.

implement classification without dimension reduction operation. By using the parameter optimization function in LibSVM, we can automatically find the most suitable c and g; finally, the value of c is 2048 and g is 0.5. The classification accuracy of train set is 66.84% by Weka. For a total of 4428 samples, 653 positive samples and 815 negative samples are misclassified. For test set, the accuracy reaches 71.77%. For 1832 samples, there are 94 positive samples and 416 negative samples that are misclassified.

Simultaneously, we also test the datasets, which are processed by dimension reduction to judge the effect of MRMD method dimension reduction on classification results. First, we use MRMD for training set. After dimension reduction, the sample space dimension is reduced from 39 to 21. Second, we leave the feature of the test set selected by MRMD in training set and delete the others. Then, we do the same operation for the reduced dimension dataset. The optimal parameters c and g are 128 and 2, respectively. The classified accuracy of training set is 66.96% and test set is 72.16%. In train set, 656 positive samples and 807 negative samples are misclassified. In test set, 94 positive samples and 416 negative samples are misclassified.

To show more vividly the number of samples that have not been dimensionally reduced and have been correctly predicted, we have drawn Figure 2. *TP* represents the vesicle transporters predicted correctly, and *TN* is regarded as the nonvesicular transporters predicted incorrectly. From Figure 2, we can clearly see that the number of correctly predicted samples after dimensionality reduction is basically the same as that without dimensionality reduction. However, from another point of view, although this technique cannot classify more samples, correctly, it eliminates some features that do not contribute much to the classification and reduces the complexity of the classification process.

Of course, if it is unreasonable and incomplete to judge the prediction effect only by the accuracy rate, we need to know other indicators about the classification results to evaluate the result more objectively. For this reason, we list the four indexes, recall, precision, accuracy, and MCC, in the performance of classification of reduced dimension and not reduced dimension and create Table 2 to represent it.

Table 2: Comparison of classification results.

	Recall	Precision	Accuracy	MCC
39 characteristics	0.718	68.65%	71.77%	0.327
21 characteristics	0.722	70.53%	72.16%	0.342

In Table 2, it is obvious that the prediction results using the 21 features after dimensionality reduction have not decreased. This proves that MRMD has no negative effect on the prediction. In addition, because MRMD calculates the contribution of each feature to classification and sorts them in the process of dimensionality reduction, we can understand which features have great differences between vesicular transporters and nonvesicular transporters. For example, according to the calculation of MRMD, the 32nd feature, called charge. G2, is ranked first after dimensionality reduction, which indicates that this feature has the greatest difference between positive and negative samples. The 13th feature, the hydrophobicity_CASG920101.G1, is in second place, which means that the degree of difference between two categories is second only to the 32nd feature and so on. The specific meaning of these characteristics can be found in chapter 2.2 of Tomii et al.: they represent different states of physical and chemical properties, such as hydrophobicity, normalized van der Waals volume, polarization, and polarizability. This partly explains whether a protein becomes a vesicle transporter because some amino acid combinations in their sequences appear physical and possess chemical properties that other proteins do not. Certainly, these are not the only factors that determine protein function.

4. Conclusion

At present, in protein classification, scholars often extract a large number of features or the classification methods used are very complex. In our research, we used CTDC feature extraction combined with MRMD feature screening and

dimensionality reduction. It is worth mentioning that the MRMD adopted to reduce the dimension, which not only reduces the number of features used in classification, but also has no negative interference to the prediction effect. Finally, we used only 21 features to complete the prediction of vesicle transporters and achieved a satisfactory result. The accuracy of our prediction method is 66% for training set by 5-fold cross validation and 72% for test set after dimension reduction. Compared with the widely used convolution neural network (CNN) or deep neural network (DNN), although it will obtain higher accuracy, there are also problems of over fitting and poor interpretability of classification process. The operation process of these methods cannot be explained, and each parameter in the classifier is adjusted by negative feedback according to the actual and theoretical results. The prediction process relies on the mutual accumulation of input and output of a series of individual neurons. It is difficult to say whether the result is related to the specific amino acid arrangement or some specific groups. However, for these classical characteristics, the sequence feature often means that there are some rules in the arrangement of amino acids in the sequence. It may be helpful for scholars to judge whether an unknown protein is a vesicular transporter. Through our study, the difference degree of each feature between positive and negative samples differs according to the calculation of MRMD. The features, like charge. G2 and hydrophobicity_Casg920101.G1, ranked first and second, respectively, and indicate that these physicochemical properties play a key role in the recognition of vesicle transporters. Moreover, the best classification results can be obtained by selecting the first 21 features, which also indicates that the content of amino acid combinations of the remaining 18 features represented between vesicular transporter and nonvesicular transporter is not significantly different. The difference in the content of these groups with specific physicochemical properties also helps to explain why proteins exhibit specific functions.

Data Availability

Experimental data can be obtained from https://github.com/taozhy/identifying-vesicle-transport-proteins or ask the author directly by email: 1765145064@qq.com.

Conflicts of Interest

The authors have declared no competing interests.

Authors' Contributions

Yuming Zhao and Yanjuan Li conceived and designed the project. Zhiyu Tao and Benzhi Dong conducted experiments and analyzed the data. Zhiyu Tao and Yanjuan Li wrote the paper. Zhixia Teng and Yuming Zhao revised the manuscript. All authors read and approved the final manuscript. Zhiyu Tao and Yanjuan Li equally contributed equally to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (grant numbers 61971117 and 61901103) and in part by the Natural Science Foundation of Heilongjiang Province (grant number LH2019F002).

References

[1] Y.-J. Tang, Y.-H. Pang, and B. Liu, "IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning," *Bioinformatics*, 2020.

[2] Y. Xu, W. Zhao, S. D. Olson, K. S. Prabhakara, and X. Zhou, "Alternative splicing links histone modifications to stem cell fate decision," *Genome Biology*, vol. 19, no. 1, 2018.

[3] Y. Xu, Y. Wang, J. Luo, W. Zhao, and X. Zhou, "Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision," *Nucleic Acids Research*, vol. 45, no. 21, pp. 12100–12112, 2017.

[4] T. D. Capellini, G. Vaccari, E. Ferretti et al., "Scapula development is governed by genetic interactions of *Pbx1* with its family members and with *Emx2* via their cooperative control of *Alx1*," *Development*, vol. 137, no. 15, pp. 2559–2569, 2010.

[5] J. Shao, K. Yan, and B. Liu, "FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network," *Briefings in bioinformatics*, 2020.

[6] L. Xu, G. Liang, S. Shi, and C. Liao, "SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins," *International Journal of Molecular Sciences*, vol. 19, no. 6, 2018.

[7] L. Yu, F. Xu, and L. Gao, "Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.

[8] Q. Yang, Y. Wang, Y. Zhang et al., "NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data," *Nucleic Acids Research*, vol. 48, no. W1, pp. W436–W448, 2020.

[9] J. Yin, W. Sun, F. Li et al., "VARIDT 1.0: variability of drug transporter database," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1042–D1050, 2020.

[10] Y. Wang, S. Zhang, F. Li et al., "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1031–D1041, 2020.

[11] 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.

[12] M. L. Liu, W. Su, Z. X. Guan et al., "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein & Peptide Science*, vol. 21, 2020.

[13] S. H. Li, J. Zhang, Y. W. Zhao et al., "iPhoPred: a predictor for identifying phosphorylation sites in human protein," *IEEE Access*, vol. 7, pp. 177517–177528, 2019.

[14] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers in Genetics*, vol. 9, 2019.

[15] Y. H. Li, X. X. Li, J. J. Hong et al., "Clinical trials, progression-speed differentiating features and swiftness rule of the

innovative targets of first-in-class drugs," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 649–662, 2020.

[16] J. Tang, J. Fu, Y. Wang et al., "ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 621–636, 2020.

[17] Q. Yang, B. Li, J. Tang et al., "Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 1058–1068, 2020.

[18] J. Tang, Y. Wang, J. Fu et al., "A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1378–1390, 2020.

[19] L. Xu, G. Liang, B. Chen, X. Tan, H. Xiang, and C. Liao, "A computational method for the identification of endolysins and autolysins," *Protein & Peptide Letters*, vol. 27, no. 4, pp. 329–336, 2020.

[20] C. Meng, F. Guo, and Q. Zou, "CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes," *Computational Biology and Chemistry*, vol. 87, article 107304, 2020.

[21] 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.

[22] J. Hong, Y. Luo, Y. Zhang et al., "Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1437–1447, 2020.

[23] F. Li, Y. Zhou, X. Zhang et al., "SSizer: determining the sample sufficiency for comparative biological study," *Journal of Molecular Biology*, vol. 432, no. 11, pp. 3411–3421, 2020.

[24] J. E. Rothman and L. Orci, "Movement of proteins through the Golgi stack: a molecular dissection of vesicular transport," *The FASEB Journal*, vol. 4, no. 5, pp. 1460–1468, 1990.

[25] Y. Liu and and R. H. Edwards, "The role of vesicular transport proteins in synaptic transmission and neural degeneration," *Annual Review of Neuroscience*, vol. 20, no. 1, pp. 125–156, 1997.

[26] J. J. Rilstone, R. A. Alkhater, and B. A. Minassian, "Brain dopamine–serotonin vesicular transport disease and its treatment," *New England Journal of Medicine*, vol. 368, no. 6, pp. 543–550, 2013.

[27] N. Q. K. Le, E. K. Y. Yapp, N. Nagasundaram, M. C. H. Chua, and H.-Y. Yeh, "Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 1245–1254, 2019.

[28] 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, 2019.

[29] K. Yan, J. Wen, Y. Xu, and B. Liu, "Protein fold recognition based on auto-weighted multi-view graph embedding learning model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[30] 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, no. 3, pp. 234–240, 2019.

[31] W. Chen, P. Feng, and F. Nie, "iATP: a sequence based method for identifying anti-tubercular peptides," *Medicinal Chemistry*, vol. 16, no. 5, 2020.

[32] Q. X. Yang, Y. X. Wang, F. C. Li et al., "Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility," *CNS Neuroscience & Therapeutics*, vol. 25, no. 9, pp. 1054–1063, 2019.

[33] J. Tang, J. Fu, Y. Wang et al., "Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains," *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. 1683–1699, 2019.

[34] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, "iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 412–420, 2018.

[35] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation," *Bioinformatics*, vol. 35, no. 16, pp. 2757–2765, 2019.

[36] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers in Pharmacology*, vol. 9, 2018.

[37] 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, 2020.

[38] 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.

[39] 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.

[40] K. Tomii and M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," *Protein Engineering, Design and Selection*, vol. 9, no. 1, pp. 27–36, 1996.

[41] 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.

[42] 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.

[43] H. Yang, W. Yang, F. Y. Dao et al., "A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae," *Briefings in Bioinformatics*, 2019.

[44] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, 2019.

[45] 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, 2018.

[46] 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.

[47] L. Yu and L. Gao, "Human pathway-based disease network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1240–1249, 2019.

[48] B. Li, J. Tang, Q. Yang et al., "NOREVA: normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W162–W170, 2017.

[49] Y. H. Li, C. Y. Yu, X. X. Li et al., "Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1121–D1127, 2018.

[50] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, "SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome," *Molecular Therapy-Nucleic Acids*, vol. 18, pp. 131–141, 2019.

[51] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation," *Molecular Therapy-Nucleic Acids*, vol. 16, pp. 733–744, 2019.

[52] B. Manavalan, T. H. Shin, and G. Lee, "PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine," *Frontiers in Microbiology*, vol. 9, 2018.

[53] Q. H. Jiang, G. Wang, S. Jin, Y. Li, and Y. 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.

[54] 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.

[55] L. Wei, P. Xing, G. Shi, Z. Ji, and Q. Zou, "Fast prediction of methylation sites using sequence-based feature selection technique," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1264–1273, 2019.

[56] 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.

[57] L. Yu, J. Zhao, and L. Gao, "Predicting potential drugs for breast cancer based on miRNA and tissue specificity," *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 971–982, 2018.

[58] 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, article 110098, 2020.

[59] W. Xue, F. Yang, P. Wang et al., "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 Chemical Neuroscience*, vol. 9, no. 5, pp. 1128–1140, 2018.

[60] J. Fu, J. Tang, Y. Wang et al., "Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification," *Frontiers in Pharmacology*, vol. 9, 2018.

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

[62] Q. Zou, J. Zeng, L. Cao, and R. Ji, Part 2, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[63] M. Niu, J. Zhang, Y. Li et al., "CirRNAPL: a web server for the identification of circRNA based on extreme learning machine," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 834–842, 2020.

[64] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt independence criterion," *Neurocomputing*, vol. 383, pp. 257–269, 2020.

[65] 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.

[66] R. Su, X. Liu, and L. Wei, "MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 687–698, 2020.

[67] 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.

[68] K. Yan, X. Fang, Y. Xu, and B. Liu, "Protein fold recognition based on multi-view modeling," *Bioinformatics*, vol. 35, no. 17, pp. 2982–2990, 2019.

[69] 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.

[70] H. Lv, F. Y. Dao, D. Zhang et al., "iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes," *iScience*, vol. 23, no. 4, article 100991, 2020.

[71] K. Liu and W. Chen, "iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, 2020.

[72] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, "Is there any sequence feature in the RNA pseudouridine modification prediction problem?," *Molecular Therapy-Nucleic Acids*, vol. 19, pp. 293–303, 2020.

[73] L. Xu, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "k-Skip-n-gram-RF: a random forest based method for Alzheimer's disease protein identification," *Frontiers in Genetics*, vol. 10, 2019.

[74] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening," *Medicinal Research Reviews*, vol. 40, no. 4, pp. 1276–1314, 2020.

[75] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, and D. C. Yang, "mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides," *International Journal of Molecular Sciences*, vol. 20, no. 8, 2019.

[76] M. Hasan, B. Manavalan, S. Khatun, and H. Kurata, "i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome," *International Journal of Biological Macromolecules*, vol. 157, pp. 752–758, 2020.

[77] 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 γ-stimulated HeLa cells," *PLoS One*, vol. 5, no. 7, article e11794, 2010.

[78] 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, 2008.

[79] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A multimodal deep learning framework for predicting drug-drug interaction events," *Bioinformatics*, 2020.

[80] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.

[81] W. Zhang, K. Jing, F. Huang et al., "SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions," *Information Sciences*, vol. 497, pp. 189–201, 2019.

[82] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLoS Computational Biology*, vol. 14, no. 12, article e1006616, 2018.

[83] 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.

[84] 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.

[85] J. Wang, S. Chen, L. Dong, and G. Wang, "CHTKC: a robust and efficient k-mer counting algorithm based on a lock-free chaining hash table," *Briefings in Bioinformatics*, 2020.