Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 8845133, 12 pages
https://doi.org/10.1155/2020/8845133

Research Article

PredAmyl-MLP: Prediction of Amyloid Proteins Using Multilayer Perceptron

Yanjuan Li[1], **Zitong Zhang**[1], **Zhixia Teng**[1], **and Xiaoyan Liu**[2]

[1] *College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China*
[2] *College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150040, China*

Correspondence should be addressed to Zhixia Teng; tengzhixia@nefu.edu.cn

Received 14 September 2020; Revised 6 October 2020; Accepted 31 October 2020; Published 21 November 2020

Academic Editor: Hui Ding

Amyloid is generally an aggregate of insoluble fibrin; its abnormal deposition is the pathogenic mechanism of various diseases, such as Alzheimer's disease and type II diabetes. Therefore, accurately identifying amyloid is necessary to understand its role in pathology. We proposed a machine learning-based prediction model called PredAmyl-MLP, which consists of the following three steps: feature extraction, feature selection, and classification. In the step of feature extraction, seven feature extraction algorithms and different combinations of them are investigated, and the combination of SVMProt-188D and tripeptide composition (TPC) is selected according to the experimental results. In the step of feature selection, maximum relevant maximum distance (MRMD) and binomial distribution (BD) are, respectively, used to remove the redundant or noise features, and the appropriate features are selected according to the experimental results. In the step of classification, we employed multilayer perceptron (MLP) to train the prediction model. The 10-fold cross-validation results show that the overall accuracy of PredAmyl-MLP reached 91.59%, and the performance was better than the existing methods.

1. Introduction

Amyloid is an insoluble fibrous protein formed by the aggregation of certain misfolded proteins [1]. They are found in bacteria, fungi, yeast, and mammals [2]; the diversity of functions is comparable to soluble proteins. Amyloid proteins play an important role in the formation of biofilms [3], the binding and storage of peptide hormones [4], antimicrobial activity [5], and the antiviral innate immune response [6]. But not all amyloid proteins are beneficial, the extracellular deposition of amyloid fibrils can cause a series of diseases such as Alzheimer's diseases [7], type II diabetes, and Parkinson's disease [8, 9]. To understand amyloid proteins and related diseases deeply, researchers have carried out a lot of work on amyloid proteins, including amyloidosis [10, 11], polymorphs of amyloid proteins at the molecular level [12], amyloid region [13], and antibody amyloid [14].

Studies have shown that not all regions of polypeptides contribute equally to its aggregation; only some short specific amino acid sequences can act as facilitators of amyloid fibril formation [15, 16]. Therefore, many computational methods for detecting the amyloid-forming regions have been proposed. AGGRESCAN [17] is a web tool, which identifies the aggregation-prone regions in the sequence based on the intrinsic aggregation-prone profile of amino acids and their relative positions. Due to its dependence on the analysis of linear sequences, it is difficult for AGGRESCAN to predict the aggregation properties of folded proteins. Zambrano et al. improve AGGRESCAN and propose a new method called AGGRESCAN3D (A3D for short) [18]. By using many factors affecting protein aggregation, A3D obtains a more accurate prediction for globular proteins. Zyggregator [19] predicts the aggregation-prone regions of polypeptides based on the physical and chemical properties of protein primary structure, such as hydrophobicity and secondary structure tendency. Based on the formation mechanism of β-sheets in amyloid aggregates, PASTA [20] uses the energy function to calculate the amino acid fragments in the sequence. FoldAmyloid [21] introduces the expected probability of hydrogen bonds and the packing density of residues to detect the

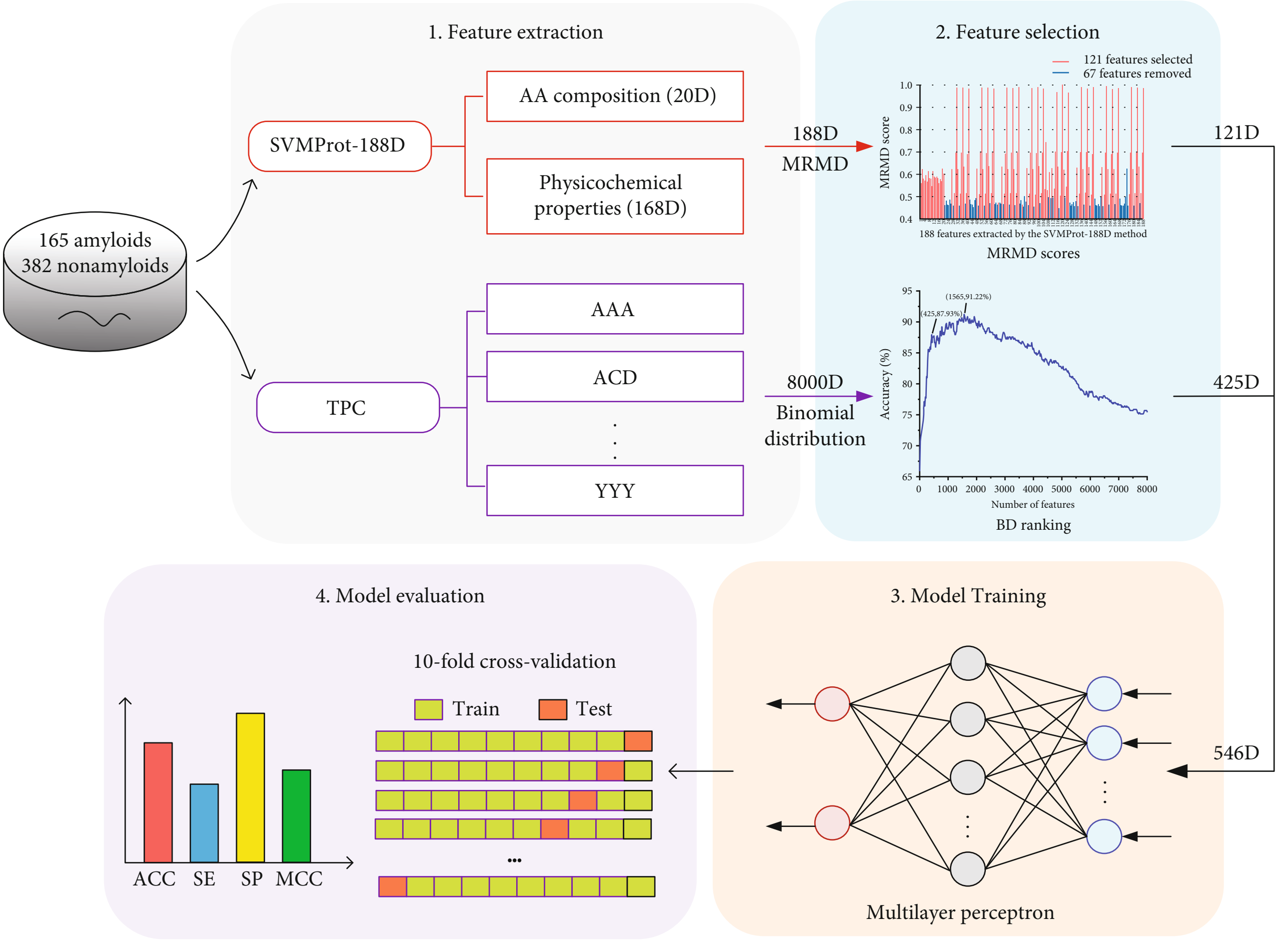

Figure 1: The frame chart of the PredAmyl-MLP predictor.

amyloidogenic regions in polypeptide chains. Maurer-Stroh's method [22] is a prediction algorithm using position-specific scoring matrices to determine amyloid formation sequences.

The prediction principles of the above methods are different and have their own advantages and disadvantages. The idea of combining different predictors to improve the identification ability was first introduced in AmylPred [23], subsequently followed by the improved version AmylPred2 [24]. AmylPred2 combines 11 different individual predictors to form a consensus prediction of the amyloidogenic region. The consensus of AmylPred2 is based on binary predictions; Emily et al. improves the weighting process and proposes MetAmyl [25]. MetAmyl introduces the meta-prediction whose input is the prediction scores of base-prediction based on a statistical approach.

In recent years, machine learning has increasingly become a favorite tool in the field of bioinformatics [26–35]. Many scholars try to use machine learning algorithms to predict amyloidogenic propensity. PASTA 2.0 [36] not only uses a pairwise energy potential to predict amyloid fibril regions but also uses machine learning algorithms to detect secondary structure. FISH Amyloid [37] proposes an original machine learning classification method to investigate co-occurrence patterns in the sequence based on the assumption that the distribution of residues in amyloid-forming fragments is position-specific. APPNN [38] is a phenomenological amyloid formation propensity predictor established on recursive feature selection and feed-forward neural network. Experimental results show that APPNN has a high accuracy value compared with other amyloidogenic propensity prediction methods.

These methods can help us understand amyloid-related diseases and find potential therapeutic targets. However, their work focuses on predicting the amyloid-forming region of a given sequence, rather than identifying whether this sequence is amyloid. Niu et al. propose RFAmyloid [39] to identify amyloid based on random forest, which obtains an accuracy of 89%. Although high accuracy has been achieved, there are still many aspects worthy of further investigation, such as redundant features due to no feature selection. In this paper, we aim to propose a new amyloid predictor, PredAmyl-MLP, to further improve the prediction performance.

2. Materials and Methods

2.1. Framework of PredAmyl-MLP. In this paper, we proposed a new amyloid predictor called PredAmyl-MLP, the framework of which is shown in Figure 1. First, we, respectively, extracted 188-dimensional vectors and 8000-dimensional vectors to represent protein sequences by using the SVMProt-

Table 1: Three groups of amino acids divided by 8 different physicochemical properties.

Physicochemical property	Class1	Class2	Class3
Hydrophobicity	RKEDQN	GASTPHY	CVLIMFW
Normalized Van der Waals volume	GASCTPD	NVEQIL	MHKFRYW
Polarity	LIFWCMVY	PATGS	HQRKNED
Polarizability	GASDT	CPNVEQIL	KMHFRYW
Charge	KR	ANCQGHILMFPSTWYV	DE
Surface tension	ILMFPWYV	KTSEC	GQDNAHR
Secondary structure	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	ALFCGIVM	RKQEND	MPSTHY

188D method and the TPC method. Next, we reduced the 188-dimensional vectors to 121-dimensional vectors using the MRMD method, reduced 8000-dimensional vectors to 425-dimensional vectors using the BD method, and then generated multi-feature vectors by combining the 121-dimensional and 425-dimensional vectors. Finally, we constructed a multilayer perceptron-based classifier that takes the multi-feature vectors as input. We will introduce the datasets, feature extraction, feature selection, and classifiers in detail in the following section.

2.2. Dataset. In this study, we utilized the dataset constructed by Niu et al. who developed a web server named RFAmyloid [39] to identify amyloid proteins. There are three reasons for considering this dataset as our experimental dataset. First, the dataset was collected from the UniProt database (http://www.uniprot.org/) and the AmyPro database (http://www.amypro.net/); thus, it is reliable. Second, the authors employed the program CD-HIT [40] to cluster proteins that meet a similarity threshold and removed redundant and homology-biased sequences [41]. Finally, and most importantly, using the same dataset allows us to compare the proposed method fairly with existing methods. The final dataset consists of 165 amyloid proteins (positive examples) and 382 non-amyloid proteins (negative examples).

2.3. Feature Extraction. The first and the most important step of designing a protein predictor is how to represent protein by features that can effectively discriminate positive samples from negative samples [42–48]. In this paper, we try to encode amyloid proteins with multi-feature, which consists of two basic feature representation methods, namely, SVMProt-188D and Tripeptide compositions (TPC). SVMProt-188D is based on the composition and physicochemical properties of amino acids. It has achieved good performance on several bioinformatics applications such as human protein subcellular localization prediction [49–52], TATA binding protein identification [53], and protein functional family prediction [54–59]. TPC is based on the tripeptide composition of protein. It also has been widely applied to solve many bioinformatics problems such as hormone binding protein identification [60], the prediction of subcellular localization of mycobacterial proteins, and the identification of cancerlectins [61–63]. In this paper, we, respectively, extract SVMProt-188D and TPC features from a protein and combine the features to represent the protein. The experimental results show that the multi-feature can effectively encode the protein, which is shown in Section 3.2. The detail of SVMProt-188D and TPC is as follows.

2.3.1. SVMProt-188D. Based on the composition and physicochemical properties of amino acids, the SVMProt-188D method encodes a protein as a 188-dimensional feature vector. The first 20 dimensions are represented by calculating the frequencies of 20 natural amino acids (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y in alphabetical order) in the sequence. The formula can be defined as

$$(V_1, V_2, \cdots, V_{20}) = \frac{N_i}{L}, \tag{1}$$

where N_i represents the number of the *ith* amino acid in the protein sequence and L represents the length of a sequence. Obviously, $\sum V_i = 1$.

The latter dimensions are correlated with eight physicochemical properties including hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility. Each property is divided into three categories, and 20 amino acids belong to different categories (listed in Table 1). All physicochemical properties are described by three descriptors C (composition), T (transition), and D (distribution). The C, T, and D descriptors of each property consist of 3, 3, and 15 numbers, respectively. C is the frequency of amino acids in a specific category. T is the percent frequency that amino acids in a category followed by amino acids in another category, such as the transitions from hydrophilic to hydrophobic or from neutral to hydrophilic. D calculates the proportions of the chain length of the first, 25, 50, 75, and 100% amino acids in a specific category and enlarges the calculations by 100 times.

Therefore, after analyzing the composition and eight physicochemical properties of amino acids, we can obtain a total of $20 + (\mathrm{C} + \mathrm{T} + \mathrm{D}) \times 8 = 188$ features.

2.3.2. TPC. The TPC method represents sequences based on the tripeptide composition of protein. Three amino acids are linked by peptide bonds to form a tripeptide, thus producing $20 \times 20 \times 20 = 8000$ possible tripeptides. TPC transforms 8000 tripeptides into an 8000-dimensional feature vector that can express a protein sequence. The formula is defined as follows:

$$F = [f_1, f_2, \cdots, f_{8000}]^T, \tag{2}$$

where T is the transposition of a vector and f_i is the frequency of the tripeptide in the sequence, which can be calculated as

$$f_i = \frac{N_i}{L-2}, \tag{3}$$

where N_i is the number of the *ith* tripeptide and L represents the length of a sequence.

2.4. Feature Selection. Feature selection plays an important role in the improvement of identification performance. It can remove redundant or noise features. We adopted the maximum relevant maximum distance (MRMD) [64] method to select optimal features from SVMProt-188D features and adopted the binomial distribution (BD) [65] method to select optimal features from TPC features. The principles of the two feature selection methods are as follows.

2.4.1. MRMD. Most dimensionality reduction methods focus on the correlation between features and target class, ignoring the redundancy of features [64]. However, the effect of highly correlated feature vectors on classification cannot be superposed. The MRMD method considers these two aspects to score features. Therefore, the score for each feature contains two components, the maximum relevant MR score and the maximum distance MD score, which can be defined as

$$\max\left(\mathrm{MR}_i + \mathrm{MD}_i\right). \tag{4}$$

The relevance between feature and target class is measured by the Pearson correlation coefficient (PCC). The formula is defined as

$$PCC\left(\vec{F_i}, \vec{C}\right) = \frac{\sum_{k=1}^{N}\left(F_{ik} - \overline{F_i}\right)\left(C_k - \bar{C}\right)}{\sqrt{\sum_{k=1}^{N}\left(F_{ik} - \overline{F_i}\right)^2}\sqrt{\sum_{k=1}^{N}\left(C_k - \bar{C}\right)^2}}, \tag{5}$$

where N is total number of samples, $\vec{F_i}$ and $\vec{C}$ consist of the *ith* dimension feature vector and the corresponding target class c in each sample, respectively; F_{ik} and C_k is the *kth* element of $\vec{F_i}$ and $\vec{C}$, respectively. If this feature contributes significantly to classification, the value of |PCC| will be large. Thus, the MR score for feature i is calculated as

$$\max \mathrm{MR}_i = \left|\mathrm{PCC}\left(\vec{F_i}, \vec{C}\right)\right|. \tag{6}$$

The correlation between features is evaluated by calculating the distance between features. In this work, Euclidean distance (ED), Cosine similarity (COS), and Tanimoto coefficient (TC) are employed as distance functions. The formulas are as follows:

$$\begin{aligned}
\mathrm{ED}_i &= \frac{\sum\sqrt{\sum_{k=1}^{M}\left(F_i - F_k\right)^2}}{M-1} \; (i \le k \le M, k \ne i),\\
\mathrm{COS}_i &= \frac{\sum F_i * F_k}{\|F_i\| * \|F_k\| * (M-1)} \; (i \le k \le M, k \ne i),\\
\mathrm{TC}_i &= \frac{\sum F_i * F_k}{\left(\|F_i\|^2 + \|F_k\|^2 - F_i * F_k\right) * (M-1)} \; (i \le k \le M, k \ne i),
\end{aligned} \tag{7}$$

and the MD score for feature i is defined as

$$\max \mathrm{MD}_i = \frac{1}{3}\left(\mathrm{ED}_i + \mathrm{COS}_i + \mathrm{TC}_i\right). \tag{8}$$

2.4.2. BD. In this work, the binomial distribution method [66–68] was applied to select the optimal subset from 8000 tripeptide features. First, we judged whether the occurrence of tripeptides in a certain kind of protein is random by calculating the probability of the *ith* tripeptide in the class j samples, like this:

$$P_{ij} = \sum_{k=n_{ij}}^{N_i} \frac{N_i!}{k!(N_i - k)!} q_j^{\,k}\left(1 - q_j\right)^{N_i - k}, \tag{9}$$

where q_i is the proportion of the number of tripeptides in class j samples to in all samples, n_{ij} and N_i are the occurrence number of the *ith* tripeptide in class j ($j = \{0, 1\}$) and all samples, respectively. A smaller P value indicates more certainty about the occurrence of tripeptides. Hence, the confidence level (CL) of the *ith* tripeptide in the class j samples can be defined as

$$\mathrm{CL}_{ij} = 1 - P_{ij}. \tag{10}$$

Obviously, each tripeptide feature has two CL values, and we will choose the larger one.

Then, the features are arranged in descending order by CL values to create a ranked list. The first feature subset contains only the first feature in the list, $D_1 = [f_1]^T$. And each new subset was produced when the next candidate feature was added to the previous subset. This process was repeated until all features in the list were added. The resulting 8000 feature subsets can be described as

$$D = [D_1, D_2, \cdots, D_{8000}]^T. \tag{11}$$

Finally, for every feature set, a prediction model was constructed. The optimal feature subset can be selected based on the maximum accuracy of 10-fold cross-validation.

2.5. Classifier. Waikato Environment for Knowledge Analysis (Weka) is a well-known machine learning and data mining software. In the platform of Weka, we can integrate our own algorithms and even use his own algorithms to implement the classification task. In this paper, we experimented with many classification algorithms based on the Weka

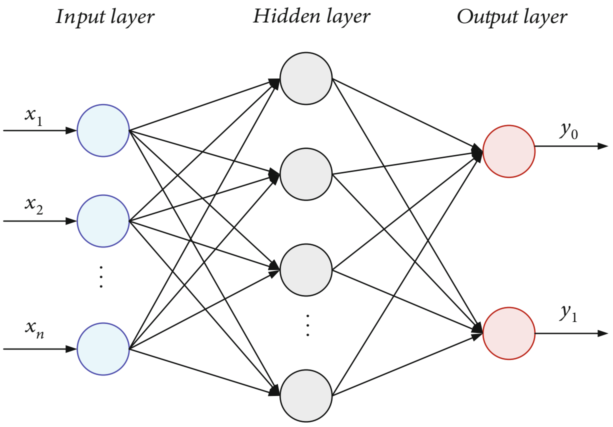

Figure 2: The structure of MLP with one hidden layer.

platform, such as random forest, naive Bayes, logistic, IBK, and bagging [69, 70]. Finally, we choose the multilayer perceptron (MLP) as our classifier, and the experimental results are shown in Section 3.3.

Artificial neural network is a machine learning algorithm that simulates the human brain. Multilayer perceptron is a kind of feedforward artificial neural network, which has a strong learning ability and robustness [71]. It performs very well in solving various practical problems and has been widely used in the field of bioinformatics, such as disease diagnosis [72, 73], the prediction of protein secondary structure [74], and gene classification [75]. MLP utilizes feature vectors as nodes in the input layer. In the training process, the output values are compared with the actual values, and error information is fed back. Based on the information, the weights continuously update until the prediction error is sufficiently small. Figure 2 is a schematic diagram of MLP. In this work, we constructed an MLP model with one hidden layer. The number of neurons in the hidden layer is set to half of the sum of the number of input features and output classes. Meanwhile, the learning rate and the number of iterations are set to 0.3 and 500, respectively.

2.6. Measurement. To evaluate the performance of our prediction model, we used four indicators commonly used in bioinformatics: accuracy (ACC), sensitivity (SE), specificity (SP), and Mathew's correlation coefficient (MCC) [76–87]. These measures are formulated as follows:

$$
\begin{aligned}
\mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},\\
\mathrm{SE} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\\
\mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}\text{-}\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},
\end{aligned}
\tag{12}
$$

where TP is the abbreviation of true positive, which means the number of amyloid proteins predicted in the positive samples; FP is the abbreviation of false positive, which means the number of amyloid proteins predicted in the negative samples; TN is the abbreviation of true negative, which means the number of non-amyloid proteins predicted in the negative samples; and FN is the abbreviation of false negative, which means the number of non-amyloid proteins predicted in the positive samples. The SE and SP, respectively, denote the predictive ability of a model in positive and negative samples. Both ACC and MCC denote the overall performance of a model. For all the indicators mentioned above, the higher scores they achieve, the better performance the models have.

3. Results and Discussion

3.1. Experiments on Feature Selection. As described in Framework of PredAmyl-MLP, we, respectively, extract SVMProt-188D and TPC features from samples and encode each sample with a multi-feature of 8188 dimensions. Training a classification model using too many feature vectors with low confidence will be relatively time-consuming, and the model may be overfitting. On the contrary, if the number of feature vectors is too small, they will not afford enough information to discriminate positive samples from negative samples. Therefore, to construct a robust and efficient prediction model, we, respectively, adopt MRMD and BD methods to choose an appropriate number of features from SVMProt-188D and TPC features. In this section, we will give the process of feature selection and experimental results.

For the 188-dimensional features extracted by the SVMProt-188D method, we assessed their importance by calculating the MRMD scores. The feature with a higher score has a more significant contribution to amyloid identification. The MRMD score consists of the Pearson correlation coefficient and distance function. The MRMD method provides three distance functions including Euclidean distance (ED), Cosine similarity (COS), and Tanimoto coefficient (TC). Different distance functions will lead to different MRMD scores for each feature. Thus, choosing an appropriate distance function is crucial for removing redundant features.

We employed support vector machines (SVM) [88, 89], a powerful classification algorithm, to examine the performance of three distance functions and select an optimal feature subset. First, we ranked the features in decreasing order of the MRMD scores to obtain the feature list. Then, we built feature subsets according to the feature order in the list. The first set contains only the feature ranking first in the list. A new set was generated when the second feature was added to the previous set. This process was repeated until all candidate features were added. Finally, the constructed 188 subsets were input into an SVM-based classifier, and the 10-fold cross-validation accuracy was obtained.

Figure 3 illustrates the performance of MRMD based on different distance functions, where MEAN represents the average of three distance function. As shown in Figure 3, ED, COS, TC, and MEAN have the best predictive

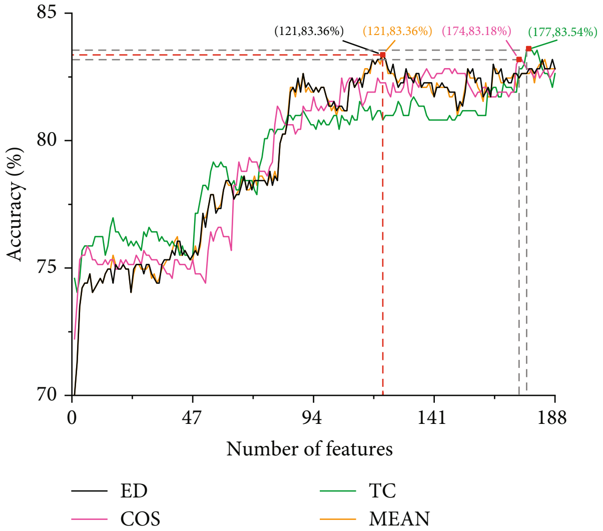

Figure 3: Comparison of different distance functions.

performance when using the top-ranked 121, 174, 177, and 121 features, respectively. Furthermore, the results obtained by ED distance function are almost identical to those obtained by the average of different distance functions. It suggests that the method using ED distance function can achieve the same effect as using the average of the three distance functions. Although the maximum accuracy of TC is slightly higher than that of ED, the number of features required for ED to obtain the best performance is much lower than that of TC. Therefore, we adopted ED as the distance function of the MRMD method and used the top 121 features in the ED ranking list to construct an optimal feature subset.

Figure 4 presents the MRMD score of each feature calculated using ED distance function, where the features marked with red are selected and the ones marked with blue are removed. As we can see from Figure 4, most of the redundant features appear continuously and concentratedly, such as 21-26, 42-47, 126-131, 147-152, and 168-175. We analyzed the reasons and found that these features were extracted based on the content of three categories of amino acids in the sequences and the transition frequency between every two categories. Such features are regarded as redundant features, possibly because they are insensitive to identifying amyloid or encodes very similar. This discovery also brings new ideas for our future research.

For the 8000 features extracted by the TPC method, we sorted them using the BD method. According to the sort order, a certain number of features are selected and formed a feature subset. Thus, we can construct 8000 feature subsets. For each subset, the SVM classifier trained with 10-fold cross-validation. The relationship between the accuracy and the number of features is shown in Figure 5. As shown in Figure 5, the accuracy reaches a maximum of 91.22% when the number of features is 1565. This number is much larger than the number of 547 samples in our dataset. The construction of a robust prediction model must take into account the time-consuming and risk of overfitting caused by high-dimensional feature vectors. Ultimately, we chose the top 425 features which can achieve an overall accuracy of 87.93% which was just slightly lower than the maximum accuracy (91.22%) produced by the top 1565 features. Therefore, the top 425 features served as the optimal feature subset in the TPC feature method.

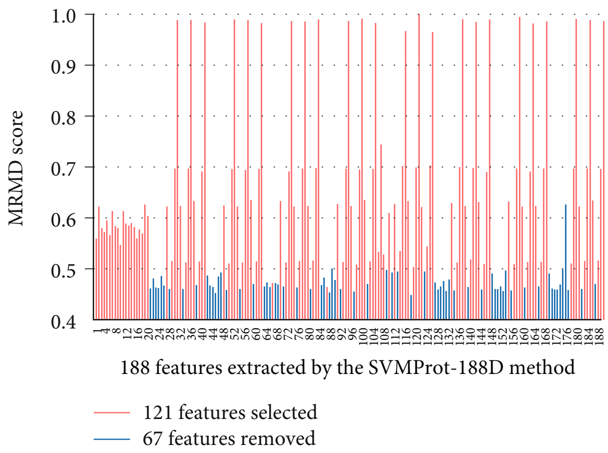

Figure 4: The MRMD scores of 188 features extracted by the SVMProt-188D method.

In summary, we, respectively, selected 121 features from SVMProt-188D features and 425 features from TPC features, then combined the 121 features and 425 features to form a multi-feature which consists of 546 features. The multi-feature is used to train the multilayer perceptron classifier in this study.

3.2. Performance of Different Features. As shown in Experiments on Feature Selection, we, respectively, extracted 188-dimensional vectors and 8000-dimensional vectors from protein sequences by using the SVMProt-188D method and the TPC method. Next, we reduced the 188-dimensional vectors to 121-dimensional vectors using the MRMD method, reduced 8000-dimensional vectors to 425-dimensional vectors using the BD method, and then generated multi-feature vectors by combining the 121-dimensional and 425-dimensional vectors. We used the multi-feature with dimensions of 546 to represent samples.

To verify the validity of the multi-feature used in this paper, we first used multilayer perceptron as the classifier and compared the multi-feature with some other features, including k-skip-2-gram [90], pseudo amino acid composition (PseAAC) [91], conjoint triad (CTriad) [92], dipeptide composition (DPC) [93], and 473D [94]. Then, three compared features with higher accuracy were combined and evaluated. Both PseAAC and DPC are based on amino acid composition. PseAAC takes into account the local information and long-range correlation of sequences. DPC represents a protein sequence through dipeptide composition information. N-gram is a common model in natural language processing, and k-skip-n-gram integrates the distance information between n residues into the traditional n-gram model. CTriad is a feature extraction method based on the neighbor relationship of amino acids. 473D encodes a sequence into a 473-dimensional feature vector based on the PSI-BLAST [95] and PSI-PRED [96] profiles.

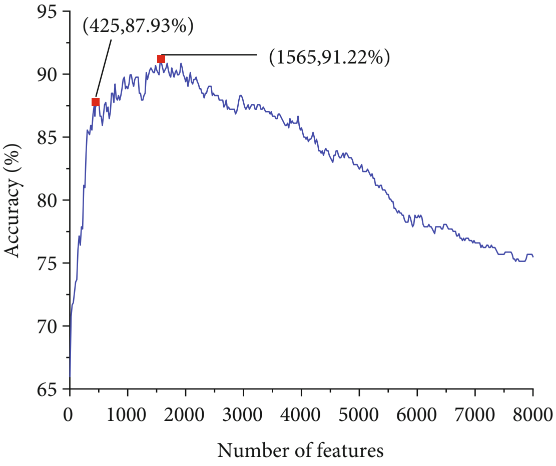

Figure 5: The accuracies of models built with different number of features.

The 10-fold cross-validation results are shown in Table 2, where both SVMProt-188 and TPC denote the final feature after feature selection. As shown in Table 2, from the indicators of ACC and MCC, the combination of SVM-188D and TPC used in this paper performs better than all other methods and has a better overall performance. According to the indicator of SE, our multi-feature also has the highest value, which demonstrates that our method performs better than other methods in identifying amyloid proteins from positive samples. According to the indicator of SP, our method is slightly lower than TPC, 473D, and the combination of CTriad and 473D. However, the values of ACC, MCC, and SE of our method are obviously higher than theirs. Especially the SE of 473D and the combination of CTriad and 473D are 0.339 and 0.036, respectively, which verify that they are biased to classify proteins as non-amyloid protein. Therefore, from the overall perspective, our method obviously performs better than all other methods.

To further illustrate that our multi-feature method has better performance regardless of the classifier, we, respectively, compared our multi-feature method with other feature extraction methods based on six different classifiers. The result is shown in Figure 6. As we can see from Figure 6, in each group of models using the same classifier, the accuracy of the combination of SVMProt 188-D and TPC is significantly higher than other feature extraction methods. Taking the classifier SGD as an example, the accuracy of the combination of SVMProt 188-D and TPC is about 9-16% higher than other methods. In general, our multi-feature method has better performance regardless of the classifier.

3.3. Performance of Different Classifiers. The selection of a classification algorithm is an important step to improve the accuracy of the model. Based on the multi-feature used in this paper, we compared multilayer perceptron with ten popular classifiers, including random forest, naïve Bayes, decision tree, AdaBoostM1, logistic, SGD, LibSVM, IBK, LWL, and bagging. SGD is a linear classifier using a stochastic gradient descent optimization algorithm. Naïve Bayes is based on Bayes' theorem and assumes that the features are independent and equally important. LibSVM is a software developed by Lin et al. to implement SVM. Logistic establishes a regression equation for the decision boundary based on the training data and classifies the test data accordingly. Decision tree divides test datasets based on the concept of entropy in informatics. AdaBoost, bagging, and random forest are ensemble classifiers. AdaBoost is an adaptive iterative algorithm, which integrates multiple weak classifiers trained on the same dataset into a strong classifier. Bagging is a parallel ensemble learning method based on bootstrap sampling. It trains a base classifier for each sampled dataset and then combines the base classifiers. Random forest is an extended variant of bagging that uses decision trees as the base classifier and introduces random attribute selection. Both IBK and LWL are lazy learning algorithms, which mean that the model is trained after receiving a test sample. IBK works by finding the k training samples nearest to a given test sample and determine the category of the given sample based on these k "neighbors," while LWL adds a concept of weighting. The results of 10-fold cross-validation are shown in Table 3.

Table 2: Comparison of different feature representation methods.

Method	ACC (%)	SE	SP	MCC
SVMProt-188D+TPC	91.59	0.836	0.950	0.798
PseAAC+473D	64.71	0.339	0.780	0.126
PseAAC+CTriad	72.76	0.491	0.830	0.333
CTriad+473D	70.56	0.036	0.995	0.119
473D+PseAAC+CTriad	67.45	0.230	0.866	0.120
SVMProt-188D	80.80	0.606	0.895	0.527
TPC	90.12	0.776	0.955	0.760
k-skip-2-gram	71.11	0.291	0.893	0.228
PseAAC	78.42	0.570	0.877	0.469
CTriad	72.57	0.345	0.890	0.281
DPC	68.37	0.345	0.830	0.193
473D	76.96	0.339	0.955	0.398

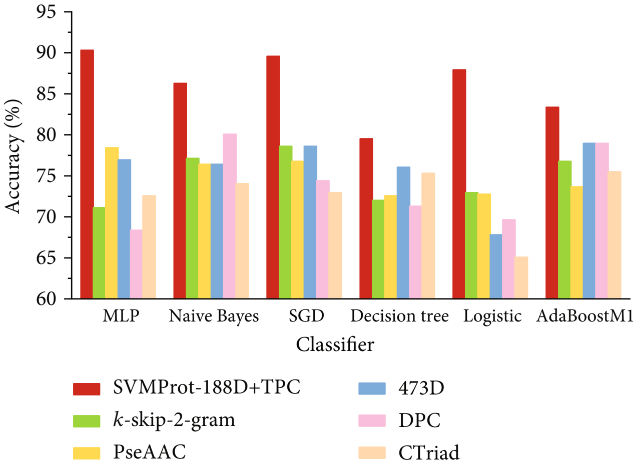

Figure 6: The accuracy of various feature extraction methods using different classifiers.

Table 3: Comparison of multilayer perceptron with other classifiers.

Method	ACC (%)	SE	SP	MCC
Multilayer perceptron	91.59	0.836	0.950	0.798
Random forest	85.00	0.642	0.940	0.629
Naïve Bayes	86.28	0.848	0.869	0.692
Decision tree	79.52	0.618	0.872	0.503
AdaBoostM1	82.81	0.612	0.921	0.574
Logistic	87.93	0.721	0.948	0.705
SGD	89.57	0.776	0.948	0.747
LibSVM	74.95	0.424	0.890	0.357
IBK	79.52	0.376	0.976	0.481
LWL	81.35	0.594	0.908	0.537
Bagging	83.36	0.588	0.940	0.585

Table 4: Comparison of our method with other existing methods.

Method	ACC (%)	SE	SP	MCC
PredAmyl-MLP	91.59	0.836	0.950	0.798
RFAmyloid	89.19	0.781	0.927	0.739
BioSeq (RF)	81.31	0.6374	0.8989	0.5626
BioSeq (SVM)	76.86	0.4953	0.9006	0.4419

In Table 3, although the multilayer perceptron method presented in this paper is slightly lower than IBK in the SP index, multilayer perceptron is obviously superior in the other three indices. In the indicator of SE, naïve Bayes achieved higher value than multilayer perceptron, but in the other three indicators of ACC, SP, and MCC, multilayer perceptron is superior to naïve Bayes. According to the indicators of ACC and MCC, multilayer perceptron is higher than all other classifiers. In general, the multilayer perceptron classifier used in this paper has better performance than other classifiers, which demonstrates that our method is effective in identifying amyloid.

3.4. Comparison with Other Methods. To further evaluate the performance of PredAmyl-MLP, we compared it with two state-of-the-art methods such as RFAmyloid [39] and BioSeq-Analysis [97] on the same dataset. BioSeq-Analysis is a platform of DNA, RNA, and protein sequence analysis that is available online at http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/PROTEIN. The SVM and random forest algorithm are used in the BioSeq-Analysis prediction method, we compared them separately. The comparison results are shown in Table 4. As we can see from Table 4, our predictor outperforms the other methods in all indicators. Furthermore, Figure 7 plots the ROC curves of the four methods. We can also see that PredAmyl-MLP is superior to existing methods in the prediction of amyloid.

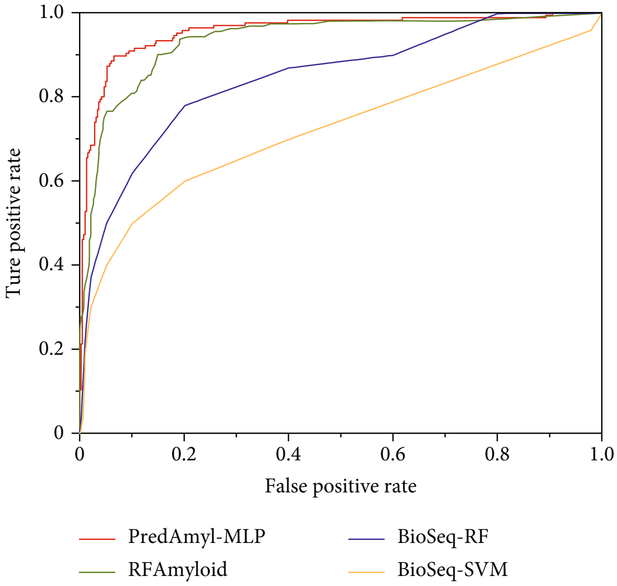

Figure 7: ROC curve for PredAmyl-MLP and other methods.

4. Conclusions

In this paper, we proposed a novel model for identifying amyloid proteins, called PredAmyl-MLP. We used the SVMProt-188D and the Tripeptide composition methods to represent protein sequences, respectively. After removing redundant features, a multilayer perception-based prediction model was constructed using mixed feature vectors. To validate the performance of PredAmyl-MLP, we compared different feature subsets, classifiers, and other methods. As a result, the features after dimension reduction can achieve better performance. Moreover, the combination of two feature representation methods significantly improves accuracy. Through a lot of experiments, PredAmyl-MLP achieved an accuracy of 91.59%, and MCC reached 0.798, outperforming other existing methods. The online server for this article is available at http://106.12.83.135:8080/amyWeb_Release/index.jsp.

In future work, we will optimize the feature representation method, using lower-dimensional feature vectors to represent amyloid sequences. Moreover, we will consider other computational intelligence models [98–102] and optimization methods [103–105] for amyloid prediction.

Data Availability

The datasets used during the present study are available from the corresponding author upon reasonable request, or can be downloaded from http://106.12.83.135:8080/amyWeb_Release/index.jsp

Conflicts of Interest

The authors declare that they have no competing interests.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (2572018BH05, 2572017CB33), National Natural Science Foundation of China (61901103, 61671189), and Natural Science Foundation of Heilongjiang Province (LH2019F002).

References

[1] C. M. Dobson, "Protein misfolding, evolution and disease," *Trends in Biochemical Sciences*, vol. 24, no. 9, pp. 329–332, 1999.

[2] D. Eisenberg and M. Jucker, "The amyloid state of proteins in human diseases," *Cell*, vol. 148, no. 6, pp. 1188–1203, 2012.

[3] P. Lembré, C. Vendrely, and P. Martino, "Identification of an amyloidogenic peptide from the Bap protein of Staphylococcus epidermidis," *Protein & Peptide Letters*, vol. 21, no. 1, pp. 75–79, 2014.

[4] S. K. Maji, M. H. Perrin, M. R. Sawaya et al., "Functional amyloids as natural storage of peptide hormones in pituitary secretory granules," vol. 325, no. 5938, pp. 328–332, 2009.

[5] S. Bieler, L. Estrada, R. Lagos, M. Baeza, J. Castilla, and C. Soto, "Amyloid formation modulates the biological activity of a bacterial protein," *Journal of Biological Chemistry*, vol. 280, no. 29, pp. 26880–26885, 2005.

[6] F. Hou, L. Sun, H. Zheng, B. Skaug, Q.-X. Jiang, and Z. J. Chen, "MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response," *Cell*, vol. 146, no. 3, pp. 448–461, 2011.

[7] C. Liu, J. Chyr, W. Zhao et al., "Genome-wide association and mechanistic studies indicate that immune response contributes to Alzheimer's disease development," *Frontiers in Genetics*, vol. 9, p. 410, 2018.

[8] L. Obici, V. Perfetti, G. Palladini, R. Moratti, and G. Merlini, "Clinical aspects of systemic amyloid diseases," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1753, no. 1, pp. 11–22, 2005.

[9] P. T. Lansbury, "Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease," *Proceedings of the National Academy of Sciences*, vol. 96, no. 7, pp. 3342–3344, 1999.

[10] J. W. D. Griffin and P. C. Bradshaw, "In silico prediction of novel residues involved in amyloid primary nucleation of human I56T and D67H lysozyme," *BMC Structural Biology*, vol. 18, no. 1, p. 9, 2018.

[11] J. Ren, C. Ren, L. Huo, F. Li, and S. Zhang, "Diffuse hepatosplenic 99mTc-pyrophosphate activity caused by amyloidosis," *Clinical Nuclear Medicine*, vol. 45, no. 3, pp. 246-247, 2020.

[12] E. L. Guenther, P. Ge, H. Trinh et al., "Atomic-level evidence for packing and positional amyloid polymorphism by segment from TDP-43 RRM2," *Nature Structural & Molecular Biology*, vol. 25, no. 4, pp. 311–319, 2018.

[13] L. C. Roisman, S. Han, M. J. Chuei, A. R. Connor, and R. Cappai, "The crystal structure of amyloid precursor-like protein 2 E2 domain completes the amyloid precursor protein family," *The FASEB Journal*, vol. 33, no. 4, pp. 5076–5081, 2019.

[14] M. P. C. David, G. P. Concepcion, and E. A. Padlan, "Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies," *BMC Bioinformatics*, vol. 11, no. 1, pp. 79–79, 2010.

[15] N. S. de Groot, I. Pallarés, F. X. Avilés, J. Vendrell, and S. Ventura, "Prediction of "hot spots" of aggregation in disease-linked polypeptides," *BMC Structural Biology*, vol. 5, no. 1, p. 18, 2005.

[16] S. Ventura, J. Zurdo, S. Narayanan et al., "Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case," *Proceedings of the National Academy of Sciences*, vol. 101, no. 19, pp. 7258–7263, 2004.

[17] O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESCAN: a server for the prediction and evaluation of \"hot spots\" of aggregation in polypeptides," *BMC Bioinformatics*, vol. 8, no. 1, p. 65, 2007.

[18] R. Zambrano, M. Jamroz, A. Szczasiuk, J. Pujols, S. Kmiecik, and S. Ventura, "AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures," *Nucleic Acids Research*, vol. 43, no. W1, pp. W306–W313, 2015.

[19] G. G. Tartaglia and M. Vendruscolo, "The Zyggregator method for predicting protein aggregation propensities," *Chemical Society Reviews*, vol. 37, no. 7, pp. 1395–1401, 2008.

[20] A. Trovato, F. Seno, and S. C. E. Tosatto, "The PASTA server for protein aggregation prediction," *Protein Engineering Design and Selection*, vol. 20, no. 10, pp. 521–523, 2007.

[21] S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya, "FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence," *Bioinformatics*, vol. 26, no. 3, pp. 326–332, 2010.

[22] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer et al., "Exploring the sequence determinants of amyloid structure using position-specific scoring matrices," *Nature Methods*, vol. 7, no. 3, pp. 237–242, 2010.

[23] K. K. Frousios, V. A. Iconomidou, C.-M. Karletidi, and S. J. Hamodrakas, "Amyloidogenic determinants are usually not buried," *BMC Structural Biology*, vol. 9, no. 1, pp. 44–44, 2009.

[24] A. C. Tsolis, N. C. Papandreou, V. A. Iconomidou, and S. J. Hamodrakas, "A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins," *PLoS ONE*, vol. 8, no. 1, p. e54175, 2013.

[25] M. Emily, A. Talvas, and C. J. P. O. Delamarche, "MetAmyl: a META-predictor for AMYLoid proteins," *PLoS ONE*, vol. 8, no. 11, p. e79722, 2013.

[26] Q. Zou, D. Mrozek, Q. Ma, and Y. Xu, "Scalable data mining algorithms in computational biology and biomedicine," *BioMed research international*, vol. 2017, 3 pages, 2017.

[27] Q. Zou, L. Chen, T. Huang, Z. Zhang, and Y. Xu, "Machine learning and graph analytics in computational biomedicine," *Artificial Intelligence in Medicine*, vol. 83, p. 1, 2017.

[28] 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.

[29] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[30] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites," *Neurocomputing*, vol. 324, pp. 3–9, 2019.

[31] L. Wei, S. Luan, L. A. E. Nagai, R. Su, and Q. Zou, "Exploring sequence-based features for the improved prediction of DNA

N4-methylcytosine sites in multiple species," *Bioinformatics*, vol. 35, no. 8, pp. 1326–1333, 2018.

[32] L. Jiang, C. Wang, J. Tang, and F. Guo, "LightCpG: a multi-view CpG sites detection on single-cell whole genome sequence data," *BMC Genomics*, vol. 20, no. 1, p. 306, 2019.

[33] Z. Zhang, J. Song, J. Tang, X. Xu, and F. Guo, "Detecting complexes from edge-weighted PPI networks via genes expression analysis," *BMC Systems Biology*, vol. 12, no. S4, p. 40, 2018.

[34] F. Guo, D. Wang, and L. Wang, "Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data," *Bioinformatics*, vol. 34, no. 12, pp. 2012–2018, 2018.

[35] Y. Liu, Y. Guo, W. Wu et al., "A machine learning-based QSAR model for benzimidazole derivatives as corrosion inhibitors by incorporating comprehensive feature selection," *Interdisciplinary Sciences*, vol. 11, no. 4, pp. 738–747, 2019.

[36] I. Walsh, F. Seno, S. C. E. Tosatto, and A. Trovato, "PASTA 2.0: an improved server for protein aggregation prediction," *Nucleic Acids Research*, vol. 42, no. W1, pp. W301–W307, 2014.

[37] P. Gasior and M. J. B. B. Kotulska, "FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids," *BMC Bioinformatics*, vol. 15, no. 1, p. 54, 2014.

[38] C. Família, S. R. Dennison, A. Quintas, and D. A. Phoenix, "Prediction of peptide and protein propensity for amyloid formation," *PLOS ONE*, vol. 10, no. 8, p. e0134679, 2015.

[39] M. Niu, Y. Li, C. Wang, and K. Han, "RFAmyloid: a web server for predicting amyloid proteins," *International Journal of Molecular Sciences*, vol. 19, no. 7, p. 2071, 2018.

[40] 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.

[41] 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.

[42] J. Zhang and B. Liu, "A review on the recent developments of sequence-based protein feature extraction methods," *Current Bioinformatics*, vol. 14, no. 3, pp. 190–199, 2019.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] J. Kang, Y. Fang, P. Yao, N. Li, Q. Tang, and J. Huang, "NeuroPP: a tool for the prediction of neuropeptide precursors based on optimal sequence composition," *Interdisciplinary Sciences*, vol. 11, no. 1, pp. 108–114, 2019.

[49] T.-H. Zhang and S.-W. Zhang, "Advances in the prediction of protein subcellular locations with machine learning," *Current Bioinformatics*, vol. 14, no. 5, pp. 406–421, 2019.

[50] 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.

[51] Y. Shen, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Critical evaluation of web-based prediction tools for human protein subcellular localization," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1628–1640, 2020.

[52] S. Wan, Y. Duan, and Q. J. P. Zou, "HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17, no. 17, article 1700262, 2017.

[53] Q. Zou, S. Wan, Y. Ju, J. Tang, and X. Zeng, "Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy," *BMC Systems Biology*, vol. 10, no. S4, p. 114, 2016.

[54] Y. H. Li, J. Y. Xu, L. Tao et al., "SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity," *PLOS ONE*, vol. 11, no. 8, article e0155290, 2016.

[55] 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.

[56] M. Naveed, M. Z. Mehboob, A. Hussain, K. Ikram, A. Talat, and N. Zeeshan, "Structural and functional annotation of conserved virulent hypothetical proteins in chlamydia trachomatis: an in-silico approach," *Current Bioinformatics*, vol. 14, no. 4, pp. 344–352, 2019.

[57] W. A. Lei, "Establishment of propagation system of mature embryo and stem segment of Fraxinus velutina," *Forestry Science & Technology*, vol. 3, p. 3, 2008.

[58] J. Li, Y. Pu, J. Tang, Q. Zou, and F. Guo, "DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 3012–3019, 2020.

[59] 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.

[60] 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.

[61] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "Discovering cancer subtypes via an accurate fusion strategy on multiple profile data," *Frontiers in Genetics*, vol. 10, 2019.

[62] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[63] H.-Y. Lai, X.-X. Chen, W. Chen, H. Tang, and H. Lin, "Sequence-based predictive modeling to identify cancerlectins," *Oncotarget*, vol. 8, no. 17, pp. 28169–28175, 2017.

[64] 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.

[65] Y. Feng and L. J. A. A. Luo, "Use of tetrapeptide signals for protein secondary-structure prediction," *Amino Acids*, vol. 35, no. 3, pp. 607–614, 2008.

[66] 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.

[67] 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*, vol. 21, 2020.

[68] 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.

[69] M. Wang, L. Yue, X. Cui et al., "Prediction of extracellular matrix proteins by fusing multiple feature information, elastic net, and random forest algorithm," *Mathematics*, vol. 8, no. 2, p. 169, 2020.

[70] 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.

[71] Y. Xu and X. Zhou, *Applications of single-cell sequencing for multiomics, in Methods in Molecular Biology*, Springer Nature, 2018.

[72] Q. Zou and Q. Ma, "The application of machine learning to disease diagnosis and treatment," *Mathematical Biosciences*, vol. 320, p. 108305, 2020.

[73] H. Işik and E. Sezer, "Diagnosis of epilepsy from electroencephalography signals using multilayer perceptron and Elman artificial neural networks and wavelet transform," *Journal of Medical Systems*, vol. 36, no. 1, pp. 1–13, 2012.

[74] X. Leo Dencelin and T. Ramkumar, "Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures," *Biomedical Research*, vol. 27, pp. 166–173, 2016.

[75] F. de Oliveira Poswar, L. C. Farias, C. A. de Carvalho Fraga et al., "Bioinformatics, interaction network analysis, and neural networks to characterize gene expression of radicular cyst and periapical granuloma," *Journal of Endodontics*, vol. 41, no. 6, pp. 877–883, 2015.

[76] 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. 21, no. 1, pp. 106–119, 2020.

[77] L. Wei, H. Chen, and R. Su, "M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning," *Molecular Therapy-Nucleic Acids*, vol. 12, pp. 635–644, 2018.

[78] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1231–1239, 2019.

[79] Z. Wang, W. He, J. Tang, and F. Guo, "Identification of highest-affinity binding sites of yeast transcription factor families," *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1876–1883, 2020.

[80] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association," *BMC Genomics*, vol. 19, no. 911, pp. 11–25, 2018.

[81] 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.

[82] 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.

[83] H. Lin, Z. Y. Liang, H. Tang, and W. Chen, "Identifying sigma70 promoters with novel pseudo nucleotide composition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1316–1321, 2019.

[84] W. Chen, F. Nie, and H. Ding, "Recent advances of computational methods for identifying bacteriophage virion proteins," *Protein and Peptide Letters*, vol. 27, no. 4, pp. 259–264, 2020.

[85] Y. Xiong, Y. Qiao, D. Kihara, H. Y. Zhang, X. Zhu, and D. Q. Wei, "Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates," *Current Drug Metabolism*, vol. 20, no. 3, pp. 229–235, 2019.

[86] 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.

[87] 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*, vol. 20, 2019.

[88] 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, p. 110098, 2020.

[89] Y. Wang, K. Liu, Q. Ma et al., "Pancreatic cancer biomarker detection by two support vector strategies for recursive feature elimination," *Biomarkers in Medicine*, vol. 13, no. 2, pp. 105–121, 2019.

[90] L. Wei, J. Tang, and Q. Zou, "SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC Genomics*, vol. 18, no. S7, p. 742, 2017.

[91] K.-. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Genetics*, vol. 43, no. 3, pp. 246–255, 2010.

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

[93] 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.

[94] L. Wei, M. Liao, X. Gao, and Q. Zou, "Enhanced protein fold prediction method through a novel feature extraction technique," *IEEE Transactions on NanoBioscience*, vol. 14, no. 6, pp. 649–659, 2015.

[95] S. F. Altschul, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[96] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[97] 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.

[98] F. G. C. Cabarle, R. T. A. de la Cruz, X. Zhang, M. Jiang, X. Liu, and X. Zeng, "On string languages generated by spiking neural P systems with structural plasticity," *IEEE Transactions on Nanobioscience*, vol. 17, no. 4, pp. 560–566, 2018.

[99] Q. Hong, R. Yan, C. Wang, and J. Sun, "Memristive circuit implementation of biological nonassociative learning mechanism and its applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 5, pp. 1036–1050, 2020.

[100] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural P systems with colored spikes," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 1106–1115, 2018.

[101] X. Zeng, S. Zhu, Y. Hou et al., "Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest," *Bioinformatics*, vol. 36, no. 9, pp. 2805–2812, 2020.

[102] B. Song, K. Li, D. Orellana-Martín, L. Valencia-Cabrera, and M. J. Pérez-Jiménez, "Cell-like P systems with evolutional symport/antiport rules and membrane creation," *Information and Computation*, vol. 270, p. 104542, 2020.

[103] H. Xu, W. Zeng, D. Zhang, and X. Zeng, "MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition," *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 517–526, 2019.

[104] Q. Hong, Z. Shi, J. Sun, and S. Du, "Memristive self-learning logic circuit with application to encoder and decoder," *Neural Computing and Applications*, vol. 32, pp. 1–13, 2020.

[105] 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*, vol. 50, no. 6, pp. 2502–2513, 2020.