Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(6): 177
https://doi.org/10.31083/j.fbl2706177

Original Research

Sequence-Based Prediction with Feature Representation Learning and Biological Function Analysis of Channel Proteins

Zheng Chen[1,2,†], Shihu Jiao[3,†], Da Zhao[1,2], Abd El-Latif Hesham[4], Quan Zou[1,3], Lei Xu[5], Mingai Sun[6,*], Lijun Zhang[2,*]

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610054 Chengdu, Sichuan, China
[2]School of Applied Chemistry and Biological Technology, Shenzhen Polytechnic, 518055 Shenzhen, Guangdong, China
[3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, 324022 Quzhou, Zhejiang, China
[4]Genetics Department, Faculty of Agriculture, Beni-Suef University, 62511 Beni-Suef, Egypt
[5]School of Electronic and Communication Engineering, Shenzhen Polytechnic, 518055 Shenzhen, Guangdong, China
[6]Department of Dentistry, Beidahuang Industry Group General Hospital, 150088 Harbin, Heilongjiang, China
*Correspondence: sunmingai6@126.com (Mingai Sun); c7zlj@szpt.edu.cn (Lijun Zhang)
[†]These authors contributed equally.
Academic Editor: Graham Pawelec
Submitted: 30 December 2021 Revised: 7 April 2022 Accepted: 19 April 2022 Published: 2 June 2022

Abstract

Background: Channel proteins are proteins that can transport molecules past the plasma membrane through free diffusion movement. Due to the cost of labor and experimental methods, developing a tool to identify channel proteins is necessary for biological research on channel proteins. **Methods**: 17 feature coding methods and four machine learning classifiers to generate 68-dimensional data probability features. Then, the two-step feature selection strategy was used to optimize the features, and the final prediction Model M16-LGBM (light gradient boosting machine) was obtained on the 16-dimensional optimal feature vector. **Results**: A new predictor, CAPs-LGBM, was proposed to identify the channel proteins effectively. **Conclusions**: CAPs-LGBM is the first channel protein machine learning predictor was used to construct the final prediction model based on protein primary sequences. The classifier performed well in the training and test sets.

Keywords: channel protein; computational prediction; light gradient boosting machine; PPI network; feature selection

1. Introduction

Channel proteins are a type of cross plasma membrane that can transport molecules of appropriate size and charged molecules from one side of the plasma membrane to the other through free diffusion motion. Channel proteins can be monomer proteins or proteins composed of multiple subunits. They are rearranged through hydrophobic amino acid chains to form aqueous channels. They do not directly interact with small charged molecules, which can diffuse freely through the aqueous channels formed by the charged hydrophilic regions of membrane proteins in lipid bilayers. The transport of channel proteins possesses a selective function, so there are various channel proteins in the cell membrane.

With high mortality, cancer is one of the most catastrophic diseases causing millions of deaths worldwide every year [1–3]. Therefore, research on the mechanism of cancer occurrence and development is still a research hotspot. However, although significant progress has been made in cancer research, there are still no good treatment strategies for cancer because the mechanism of cancer occurrence and development is too complex. Previous studies have suggested that abnormalities in channel proteins in some signaling pathways can promote the occurrence and development of cancer. For instance, chloride intracellular channel 1 (CLIC1) is a chloride channel protein. The up-regulated expression of CLIC1 is positively related to cell proliferation, invasion, migration, and angiogenesis. Chloride intracellular channel 1 promotes the progression of oral squamous cell carcinoma, and its potential mechanism may be correlated with ITGαV and ITGβ1 regulation, resulting in the activation of MAPK/ERK and MAPK/p38 signaling pathways [4]. Aquaporin-4 (AQP4) forms a heterotetramer composed of m23-AQP4 and m1-AQP4 subtypes on the plasma membrane. The isoform ratio controls the aggregation of AQP4 into the supramolecular structure, which is called the orthogonal particle array. Studies have shown that the aggregation/decomposition of AQP4 into OAP affects the biological characteristics of glioma cells, and the aggregation state of AQP4 may be an important determinant of glioma cell survival or death. The decomposition of AQP4 may enhance invasiveness, and the aggregation of AQP4 may activate the apoptotic pathway [5]. HERG (human ether - à - go related gene) K^+ current realizes essential ionic functions in the heart. HERG channels affect the migration and growth of various types of tumor cells. Studies have shown that HERG1 channel proteins take part in the growth of small cell lung cancer (SCLC) cells [6]. In

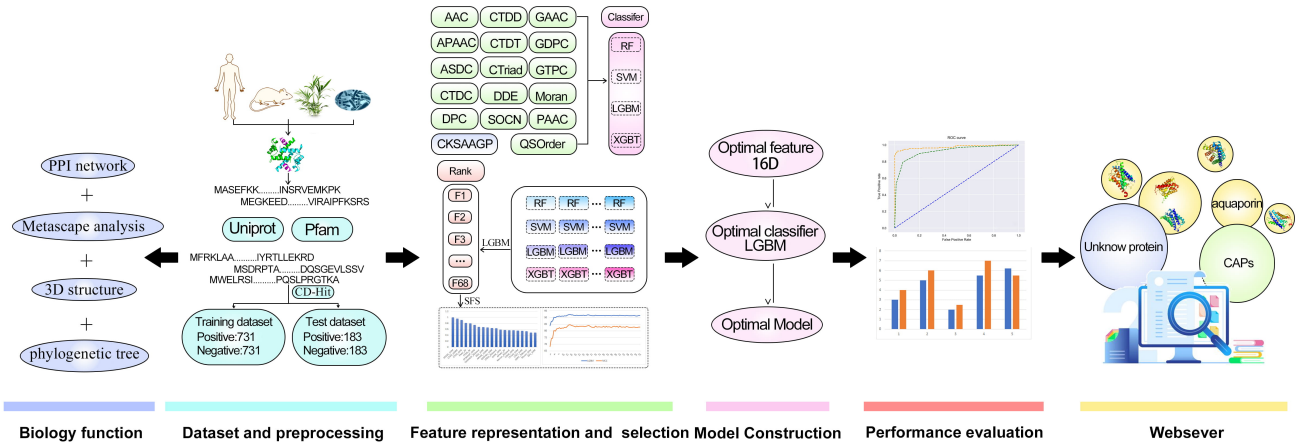

Fig. 1. The framework of CAPs-LGBM for channel protein identification.

rats and humans, voltage-gated Na^+ channels (VGSCs) are upregulated in prostate cancer (PCA), in which channel activity promotes cell invasiveness *in vitro* and metastasis *in vivo* [7].

Recent studies have shown that machine learning-based methods have been well developed, especially those related to effective feature representation algorithms [8–10]. At present, various sequences based on feature descriptors are obtained from many studies [11–13]. Taking various types of features to train classifiers is a simple method to build predictive models. A high feature dimension of integration will lead to dimension disaster, and simple integration will lead to information redundancy. One efficient way to reduce the dimension is feature selection [14]. These problems will affect the prediction performance of the model. A more effective method is necessary to use feature information. In addition, most existing feature descriptors only use sequence information to build prediction models. This may not be enough to provide enough information to accurately distinguish between real CAPs and no-CAPs. Efficient computational identification tools are good choices; however, current research efforts in this area are lacking. Due to their efficiency and convenience, machine learning-based methods have been widely used in protein function prediction [12,15–21]. Therefore, it is desirable to research reliable and effective machine learning tools for CAP identification.

In this research, Metascape was applied for the enrichment and network analysis of the biological functions of channel proteins. CAPs-LGBM is the first software model that can classify proteins as CAPs or no-CAPs. We establish the first benchmark dataset composed of 914 CAPs and 914 non-CAPs, which is publicly available to ensure the reproducibility of the proposed predictor. On this basis, we studied the feature representation learning strategy that integrates the prediction probability information into the newly derived features. To improve the prediction performance, a two-step feature optimization protocol was adopted to manually select the optimal feature subset containing 16 information features. Based on the optimal 16-dimensional probability feature, a sequence-based CAP predictor CAPs-LGBM was constructed. The results suggest that the proposed prediction model has good recognition performance. The overall framework of CAPs-LGBM is shown in Fig. 1. The prediction of channel proteins is a novel work, and there is no previous research on it. The prediction accuracy of the model has a high accuracy in our research. Meanwhile, we also optimized the prediction algorithm and developed a user-friendly online server. The model is a quick and effective way to predict whether a protein is a channel protein or not. The website is http://lab.malab.cn/~acy/CAPs-LGBM. It has the potential to promote future computing work in this field.

2. Methods

2.1 Datasets

To establish a reliable and robust CAP-LGBM predictor, a well-prepared dataset is essential. CAP and non-CAP protein sequences are composed of a positive and negative dataset for binary prediction model construction. CAP sequences were downloaded from the UniProt database (UniProtKB version 2021_03, https://www.UniProt.org/) [22]. "channel protein and reviews: Yes" was applied as the keyword to search the protein sequences. A total of 18,375 protein sequences were obtained from the UniProt database, and 2105 channel proteins were selected according to the functional annotation as the selection criteria for positive samples. Negative samples were selected from the protein family database (Pfam, version 35.0, http://pfam.xfam.org/) [23]. There are two principles for negative dataset selection: (I) each negative sample is the longest sequence from different protein families; (II) samples from positive families will be removed. Finally, a negative dataset containing CAP protein sequences was established. CD-Hit (V4.8.1) [24] was applied to remove the sequence redundancy for the positive and negative samples to avoid pro-

tein homology deviation with the threshold set to 0.8 [25]. Then, 914 CAP protein sequences were selected for machine learning. There is enough protein sequence information for the machine learning algorithm to construct an optimal model for channel proteins. Then, CAP-positive and -negative samples were retained as the channel protein dataset. The original dataset is randomly divided into two subsets at a ratio of 8:2 (Table 1), of which 80% are training sets and the rest are test sets to verify the performance of CAP. The datasets described above are freely available at http://lab.malab.cn/~acy/CAPs-LGBM.

Table 1. Sample distribution in the training and independent test datasets.

Dataset	Training	Testing
Positive	731	183
Negative	731	183

2.2 Feature Representation Learning

Sufficient feature information from the channel protein sequences is essential for accurate and reliable bioinformatics model construction [10,26–32]. Following previous research [9,11,33–38], we used a feature representation learning protocol to predict and identify the CAPs. First, 17 feature coding algorithms were used to construct the initial feature pool to represent the protein sequence. There were three categories divided by the coding methods: (1) amino acid composition characteristics features, (2) based on the characteristics of physical and chemical properties, and (3) features based on sequence order. All of the above feature descriptors are defined by ilearn tools [39]. In the second step, four common classifiers, namely, random forest (RF), XGBoost (XGBT) and SVM, were employed to train on the 17 descriptors to build the baseline prediction models. Each prediction model will provide both class information (predicted label) and probabilistic information (predicted confidence). In this work, we utilized the probabilistic information predicted by each model as a "feature". Probability information will be used as the "feature" of each model in our study, and a 68-dimensional probability feature vector will generate the prediction models (68 feature descriptors $\times$ 4 machine learning classifiers) [9].

2.3 Classifiers

Classifier choice plays an essential role in machine learning [40–42]. Various machine-learning algorithms have been applied in machine learning methods [8,43–45]. Seventeen descriptors were trained by four common classifiers: random forest (RF), XGBoost (XGBT), and SVM to construct the model. As described in previous research [46–48], all four classifiers were derived from the scikit-learn package (version 0.24). Finally, a grid search was used to adjust hyperparameters for the classifiers, and the search range is provided in **Supplementary Table 1**.

2.4 Feature Selection

In this study, a two-step feature selection method was used to improve the feature representation ability and prediction performance of the models [49]. First, the original feature set was ranked according to the classification importance score. Second, the SFS (sequential forward search) strategy was used to search the optimal feature subset from the feature list in the first step. Generally, feature selection methods are divided into packaging, filtering, and embedding methods [50,51]. The light gradient boosting machine (LGBM) is a packaging method, and the LGBM model is obtained by inputting the training data, which are sorted according to the importance score of the features. In the SFS step, additional features are obtained in the first step according to the lower to higher rank, and reconstruct the prediction model with various features. The subset with the highest accuracy of the prediction model was determined as the optimal feature set.

2.5 Performance Measurement

The validation strategies of 10-fold cross validation (CV) and testing were used to evaluate the performance of the involved models [13,52–55]. The training dataset was randomly divided into 10 subsets of approximately the same size for 10-fold CV validation. The ratio of the training data and the validation dataset was 9:1. The performance of 10 test subsets was averaged, and the result is the overall performance of the 10-fold CV test. Thus, the proposed model is verified more strictly and compared fairly with other methods.

Furthermore, accuracy (Acc), sensitivity (SE), specificity (SP), and Matthew correlation coefficient (MCC) [8,56,57] are four common metrics in binary classification tasks. Receiver operating characteristic (ROC) curves were also used to provide intuitive performance comparisons. The area under the ROC curve (AUC) was calculated and used as the main quantitative index of overall performance.

$$\begin{cases} \mathrm{SE} = \frac{TP}{TP+FN} * 100\% & (1) \\ \mathrm{SP} = \frac{TN}{TN+FP} * 100\% & (2) \\ \mathrm{Acc} = \frac{TP+TN}{TP+FP+TN+FN} * 100\% & (3) \\ \mathrm{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}} & (4) \end{cases}$$

Acc, accuracy; SE, sensitivity; SP, specificity; MCC, Matthew's correlation; TP, true positive; TN, true negative; FP, false-positive; FN, false negative.

2.6 Construction of 3D Structure and Phylogenetic Tree for CAPs

A phylogenetic tree of CAPs was constructed to analyze the evolutionary diversity of the proteins. CAP sequence alignment results were analyzed by muscle software (V3.8.1551) and used to construct a phylogenetic tree using

IQ-TREE software (multicore version 1.6.12). The best fitting model for the phylogenetic tree was VT+R6 [58]. The ultrafast bootstrap method was used for phylogenetic assessment, and 1000 replicates per method were chosen in this work [59–61]. Tree file was visualized by the iTOL website (version 6, https://itol.embl.de/). Phyre2 software (version 2.0, http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) was applied for the 3D structure prediction of CAPs. The prediction results were visualized by PyMOL (version 2.5.1, DeLano Scientific LLC) software (https://pymol.org/2/).

2.7 Metascape Analysis

Metascape (version 3.5, https://metascape.org/gp/index.html#/main/step1), as an effective tool, can analyze multiple groups of data on multiple platforms [62]. In our research, Metascape was used for enrichment analysis of CAPs. CAPs were carried out based on pathway enrichment and analysis of molecular function (MF), biological process (BP) and cell composition (CC) by the Metascape tool based on the two databases of Kyoto Encyclopedia of Genes and Genomes (KEGG, version 100.0) and Gene Ontology (GO). A network plot was constructed by a subset of enriched terms to capture the relationships between the terms. Similarity >0.3 and p values < 0.05 of the terms were connected for the edges and the 20 clusters with less than 15 terms per cluster. Based on the BioGRID, InWeb_IM, and OmniPath databases, a PPI (protein–protein interaction) network was constructed and visualized by Metascape. Cluster analysis was carried out by MCODE (minimum common oncology data elements) to identify the key clusters with default parameters in the PPI network. In addition, the significant function module selected was predicted with $p < 0.05$ significance using Metascape.

3. Results

3.1 Phylogenetic Analysis and Structural Features of CAP Proteins

The results of the phylogenetic tree (Fig. 2) showed that 134 CAP genes were divided into five groups, and the length of branches indicated the genetic relationship of CAP sequences. Among the five groups, Groups I, III, and IV contained two subfamilies. The function of AQPs is as a water molecule transport protein, which belongs to the branch of the fifth family. Calcium channel protein belongs to group IVa, potassium voltage gated channel protein belongs to groups I and III, calcium activated potassium channel subunit belongs to IVa and IIIa, sodium channel protein belongs to Ia and IVa, chloride intracellular channel protein belongs to groups Ia and IIIa, potassium/sodium channel protein belongs to IIIb and IVa. Fig. 3C also indicted the potassium/sodium channel protein involved the transportation of potassium and sodium ions. The volume-regulated anion channel subunit (leucine rich repeat containing protein) belongs to the IIa group.

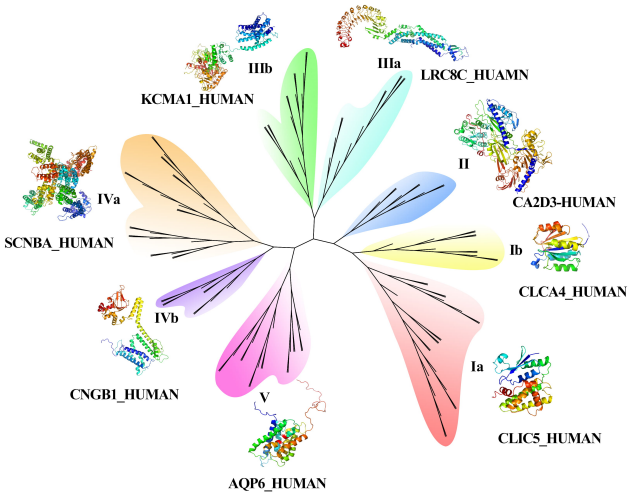

Fig. 2. Phylogenetic tree and 3D construction of CAPs.

The secondary structure of CAPs includes α-helices, β-pleated sheets, and random coils. Usually, an α-helix exists in the plasma membrane and belongs to the secondary structure of transmembrane proteins. In this study, a protein was selected from each subfamily, and the 3D structure of these proteins was constructed. The results show that all proteins had multiple α-helix proteins, indicating that the subcellular localization of most channel proteins belongs to the plasma membrane. The structure of channel proteins in Groups I and V was relatively simple, while the structure of channel proteins in groups II, III, and IV was relatively complex (Fig. 2). The composition of these structures is necessary for determining the function of channel proteins.

3.2 Detection and Enrichment Analysis of Related Genes

To identify the interaction and internal mechanism of CAP coexpressed genes, Metascape was applied to analyze the overlapping genes, enrichment, PPI network, and MCODE of CAPs and their coexpressed genes. First, the overlap analysis of coexpressed genes was divided into two methods. Specific overlapping genes were found in the coexpressed genes of CAPs of different species (Fig. 3A,B). According to the classification of channel proteins, only potassium channels and sodium channels have overlapping genes. Aquaporins and chloride channel proteins are specific molecules and ion channels, and there were no overlapping genes (Fig. 3C,D).

A total of 134 human channel proteins were screened from the positive case set for Metascape analysis. These genes were enriched based on the David database and Gene Ontology (GO). As shown in Fig. 3E, CAPs and their coexpressed genes were primarily enriched in potassium channels (R-HSA-1296071), stimuli-sensing channels (R-HSA-2672351), and regulation of membrane potential (GO:0042391).

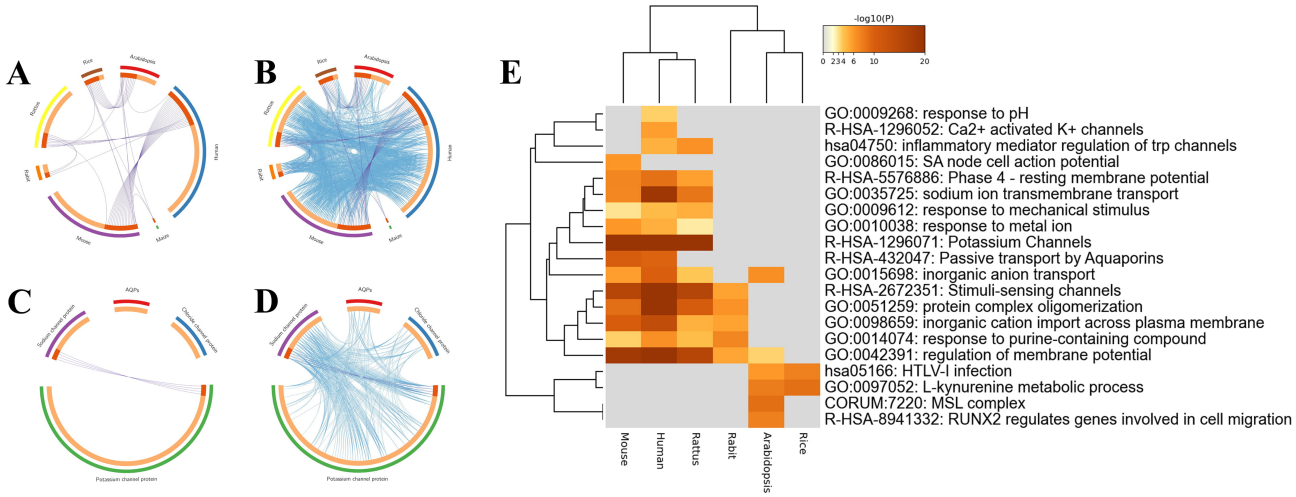

Fig. 3. Overlaps and heatmap enrichment analysis of CAPs and their coexpressed genes. (A,B) Overlap circus plot among CAPs of different species. (C,D) Overlap circus plot among CAPs of different functions. (E) Heatmap of enriched terms among CAPs.

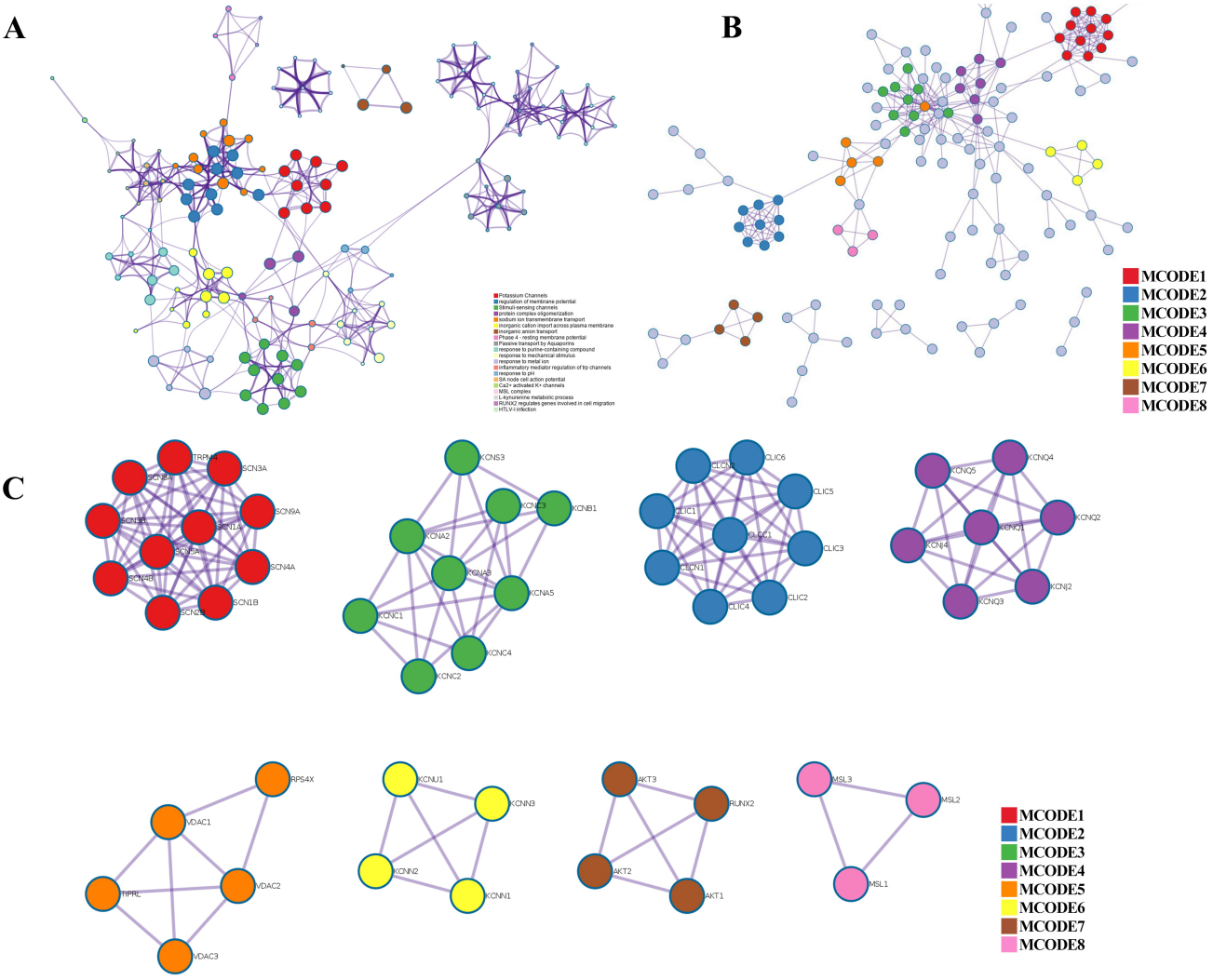

Fig. 4. PPI network, enriched terms network, and MCODE analysis of CAPs. (A) Network of the enriched terms colored by cluster ID. (B) The protein–protein interaction (PPI) network for enrichment analysis of functions and pathways. (C) MCODE components identified from the PPI network among CAPs.

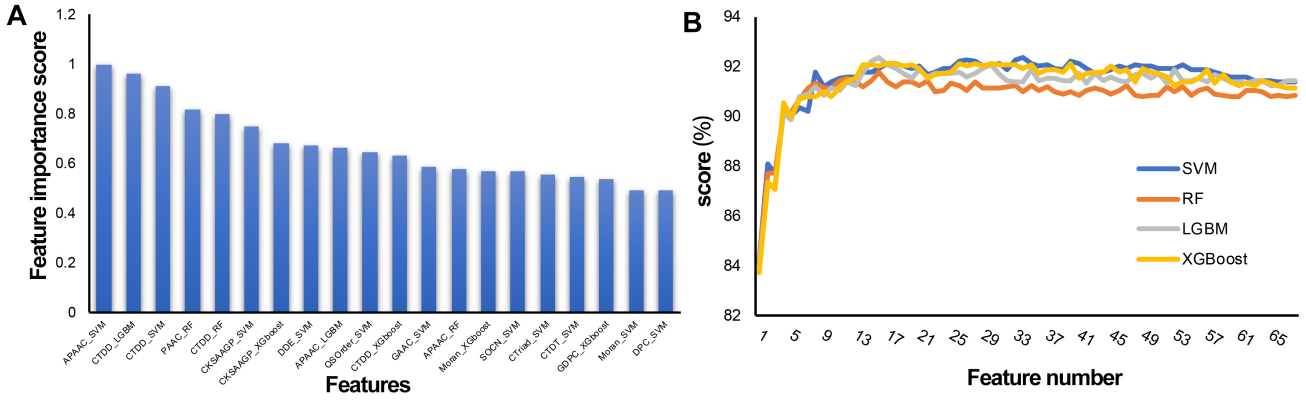

Fig. 5. Feature selection of the 68-dimensional probability feature. (A) Classification importance score of the top 20 features. (B) Tenfold cross validation accuracy on the SVM, RF, XGBoost, and LGBM classifiers with feature numbers.

3.3 Network of Enriched Terms and Screening of Hub Genes

To better understand the relationships between the terms, a network plot of enriched terms was constructed. As shown in Fig. 4A, there were 111 GO BP, 13 KEGG pathways, and 27 reaction gene sets. The data show that CAPs are enriched in potassium channels, regulation of membrane potential, stimuli sensing channels, protein complex oligomerization, and sodium ion transport.

Metascape was applied to construct the PPI network (Fig. 4B) and found important MCODE components according to the cluster score (Fig. 4C). As shown in Fig. 4C, a PPI network was successfully constructed. Eight molecular MCODE components were identified from the PPI network. MCODE 1 (SCN 1/4/5/8/9A, SCN 1/2/3/4B, TRPM4) plays an important role in the voltage-dependent sodium ion permeability of excitable membranes. MCODE 2 (CLIC 1/2/3/4/5/6, CLCN 1/2 and CLCC 1) are chloride intracellular channel proteins. MCODE 3, MCODE 4, MCODE 6, and MCODE 7 are potassium channel proteins. MCODE 5 (VDAC 1/2/3) are voltage-dependent anion-selective channel proteins. MCODE 8 (MSL 1/2/3) are mechanosensitive ion channel proteins. Importantly, these analyses applied to each MCODE component showed that the biological function was related to a series of ion transport pathways.

3.4 Feature Selection Results

To establish an effective prediction model, the probability characteristics of possible redundant information need to be processed to avoid the waste of computing resources and affect the final classification effect. In this study, a two-step feature selection method was used to identify the optimal feature representation from the original feature set. Sixty-eight features were sorted by LGBM according to classification importance. Fig. 5A shows the ranking of the top 20 features, in which the importance decreases along the x-axis. The results show that the most important feature is APAAC-SVM, indicating that it is the most effective classification among all features, followed by CTDD_LGBM and CTDD-SVM. APAAC-SVM represents the probability features derived from the APAAC descriptor on the SVM classifier. To evaluate the prediction performance of the four classifiers (SVM, RF, LGBM, and XGBoost) used in the feature representation learning scheme, 10 CV experiments and 68-dimensional classification probability features were implemented, and 68 prediction models were obtained. See **Supplementary Tables 2–5** for the results. The overall ACC performance is shown in Fig. 5B. In fact, the ACC curves of the four classifiers seem to have similar patterns. To clarify the discussion of the SFS method on the LGBM classifier, we also refer to the ACC results of the SVM, SFS, and LGBM classifiers in Fig. 5B. The results show that the SVM and LGBM classifiers are generally better than the SFS classifier. The ACC curves of the LGBM classifier increased steadily with the increase of the number of features, until the number of features was close to 16 with the Acc value of 92.34, and gradually tended towards a fluctuating plateau, this model was named M16-LGBM (Fig. 5B). The Acc curve of the SVM classifier increased steadily until it reached the maximum value of 92.34 when the feature number was 34, and the model was named M34-SVM (Fig. 5B). Considering other indicators and all the above findings, we choose the LGBM classifier to construct the predictor. We believe that only the subset of the first 16 features is optimal. To better understand the selected features, we further analyzed their composition. In fact, these 16 features are: APAAC_SVM, CTDD_LGBM, CTDD_SVM, PAAC_RF, CTDD_RF, CKSAAGP_SVM, CKSAAGP_XGboost, DDE_SVM, APAAC_LGBM, QSOrder_SVM, CTDD_XGboost, GAAC_SVM, APAAC_RF, Moran_XGboost, SOCN_SVM, and CTriad_SVM. CTDD and APAAC generate four and three features, respectively. The physicochemical information and amino acid composition information of CAPs have the strongest feature representation ability.

3.5 Comparison of the Individual Feature Descriptors and Class Features with Probability Features

To verify the effective representation ability of the probability features generated by the feature representation learning scheme, we compared the probability features M16-LGBM with class features M34-SVM and the sequence-based feature descriptors related to M16-LGBM. The 10-fold CV results of 10 sequence-based feature descriptors are summarized in **Supplementary Table 6**. We selected the M16-LGBM subset and compared them with the first three feature descriptors with the best performance, namely, DDE, QSOrder, and APAAC. Table 2 shows the results of all training sets with comparative characteristics. The overall performance of probability features was significantly improved compared with original individual features. For example, the ACC values of DDE, QSOrder, and APAAC were 88.09%, 87.34% and 88.30%, respectively, while the probability features M16-LGBM and M34-SVM had the same ACC values of 92.33%. In addition, the AUC, Sn, Sp, and MCC values of the probability features were all higher than those of the three feature descriptors. This shows that the probability information generated by the model can improve the prediction model performance. The ROC curve of comparative features is shown in Fig. 6F. We can clearly see that the AUC of probability features is the highest of all features. In conclusion, the observation results showed that the probability feature with a small dimension is more suitable for constructing our final predictor.

Table 2. 10-fold CV results for APAAC, DDE, QSOrder, M16-LGBM and M34-SVM features on the training set.

Feature	ACC (%)	AUC	Sn (%)	Sp (%)	MCC
APAAC	88.30	0.94	90.15	86.46	0.77
DDE	88.10	0.94	90.01	86.18	0.76
QSOrder	87.35	0.94	89.33	85.36	0.75
M16-LGBM	**92.34**	**0.98**	**91.93**	**92.75**	**0.85**
M34-SVM	**92.34**	0.97	92.20	92.48	**0.85**

The best performance value is highlighted in bold for clarification.

The t-distributed stochastic neighborhood embedding (t-SNE) algorithm [63] was applied to explain why our model-based probability features can effectively improve the prediction performance. As shown in Fig. 6A–C, in the feature space of APAAC, DDE, and QSOrder descriptors, many positive and negative samples are mixed together, indicating that their expression ability may not be enough to fully distinguish between real CAP and non-CAP samples. In contrast, in the two model-based feature spaces, most positive and negative samples are distributed in two significantly different clusters (Fig. 6D,E). This shows that compared with the descriptor based on individual sequences, CAPs and non-CAPs are easy to distinguish in the vector space of the probability feature model. Using our probability feature, most of the positive and negative samples are clearly separated, and only a few samples overlap in the middle region. This once again confirms that the probability feature is more effective and can reveal the potential difference pattern between CAPs and non-CAPs to improve the prediction performance.

3.6 Comparison Results of Different Ensemble Learning Methods

To verify the effectiveness of the feature representation learning method used in this paper, we explored different integrated learning schemes, including hard voting, soft voting, and stacking. We compared our CAPs-LGBM model building method with three traditional ensemble methods and evaluated their performance with 10-fold CV. Table 3 summarizes the performance results. Our CAPs-LGBM model building method was optimal in the training dataset, and similar results were found in the test dataset except for the Sn parameter. We observed that the proposed feature representation learning method CAPs-LGBM was significantly better than the traditional ensemble method. The results in Table 3 indicate that the prediction performance is significantly improved compared with other models (**Supplementary Table 6**). The feature representation learning method has the best overall performance among all comparison strategies. The ACC value on our CAP training set was 92.34%, which is 2.32%, 2.25%, and 3.69% higher than that of hard voting, soft voting, and stacking, respectively (Fig. 7A). To verify the robustness and practical applicability of CAPs-LGBM, we further compared these methods through independent testing. Similar to the training dataset, performance improvement was also observed on the test dataset. Compared with the other three ensemble predictors, the average performance of CAPs-LGBM on ACC and MCC is improved by 1.55% and 3.02%, respectively (Fig. 7B). In conclusion, our method provides satisfactory prediction results. This result shows that compared with other ensemble learning methods, feature representation learning in CAPs can make more effective use of the output of a single baseline model and help to distinguish CAPs from non-CAPs.

3.7 Case Study

In order to test whether our CAPs-LGBM toolkit can predict the protein sequence with unknown function in practical scenarios, two common proteins from UniProt database are downloaded and are using our CAPs-LGBM method for practical application prediction. The two common proteins are ubiquitin and actin, none of which are channel proteins. The sequences of these two proteins are input into our prediction toolkit CAPs-LGBM. The results showed that there are 7668 sequences of actin protein, and 701 sequences were predicted incorrectly, with an error rate of 9.1%, while among 19,757 ubiquitin sequences, 2005

Table 3. Comparison results of CAPs-LGBM with traditional ensemble learning methods.

Tools	Training dataset				Testing dataset			
	Acc (%)	Sn (%)	SP (%)	MCC	Acc (%)	Sn (%)	SP (%)	MCC
Hard voting	90.01	88.51	91.52	0.80	91.26	90.71	91.80	0.83
Soft voting	90.08	90.15	90.01	0.80	91.53	92.90	90.16	0.83
Stacking	88.65	88.92	88.37	0.77	91.26	**95.08**	87.43	0.83
CAPs-LGBM	**92.34**	**91.93**	**92.75**	**0.85**	**92.90**	93.99	**91.80**	**0.86**

The best performance value is highlighted in bold for clarification.

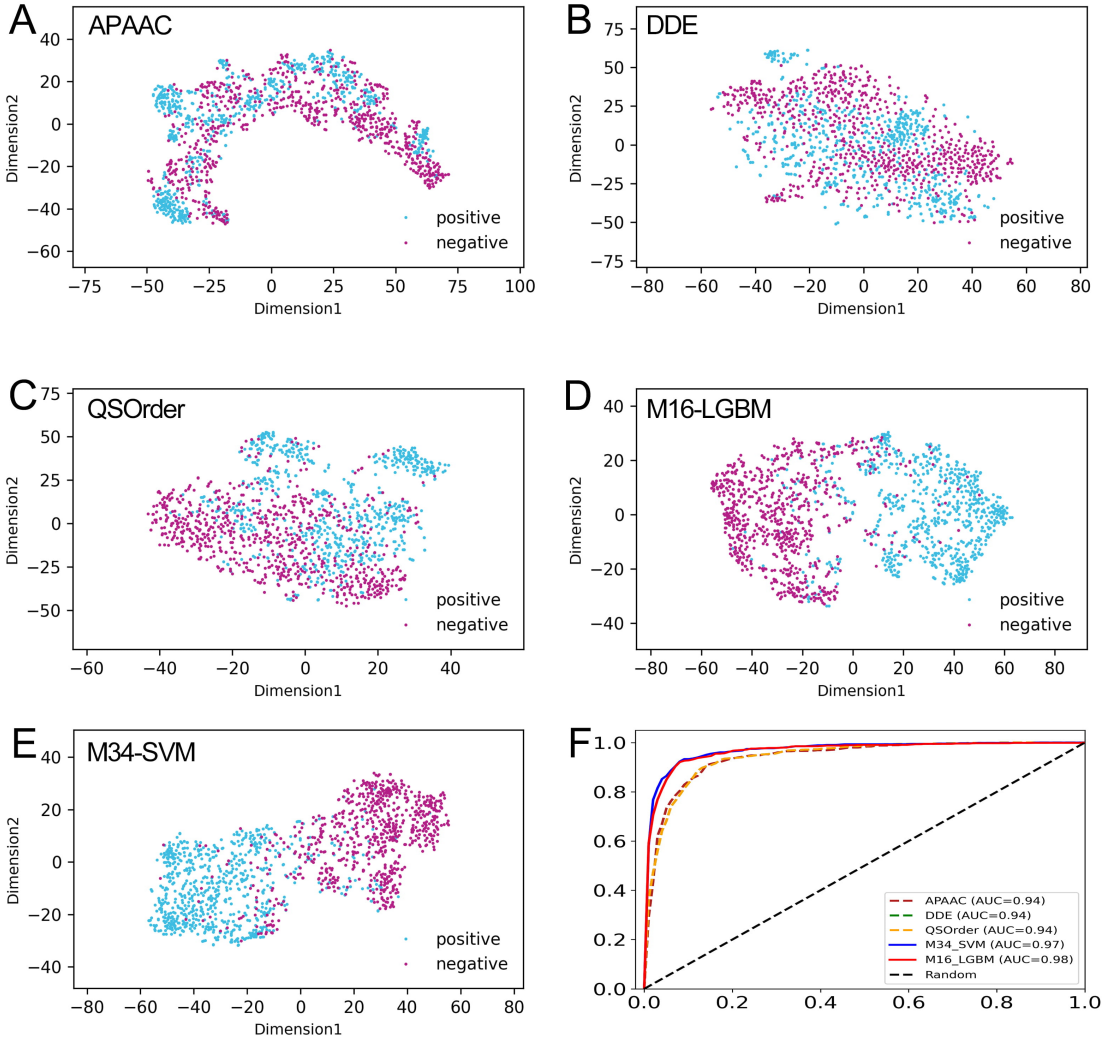

Fig. 6. Comparison of our optimal feature with the class feature and individual feature descriptors. (A–E) The t-SNE distribution of APAAC, DDE, QSOrder, M16-LGBM, and M34-SVM features. (F) The ROC curves of the APAAC, DDE, QSOrder, M16-LGBM, and M34-SVM features on the training set.

sequences were predicted incorrectly, with a prediction error rate of 10.1%. Therefore, our prediction toolkit CAPs-LGBM can obtain accurate results relatively in channel protein prediction. The results of these sequences are particularly meaningful for further experimental validation.

3.8 Web Server Implementation

To facilitate the identification of CAPs by researchers, we built a user-friendly online web server named CAPs-LGBM, which is freely available at http://lab.malab.cn/~acy/CAPs-LGBM. To validate our findings, the benchmark dataset has been applied on the online server. A simple guideline was obtained to provide researchers on the use method for the CAPs-LGBM webserver. First, users need to put the query sequence in fasta format in the left input box and click the submit button. Finally, the results are displayed on the right result box. To restart a new task, a clear button or the resubmit button should be selected to clear the

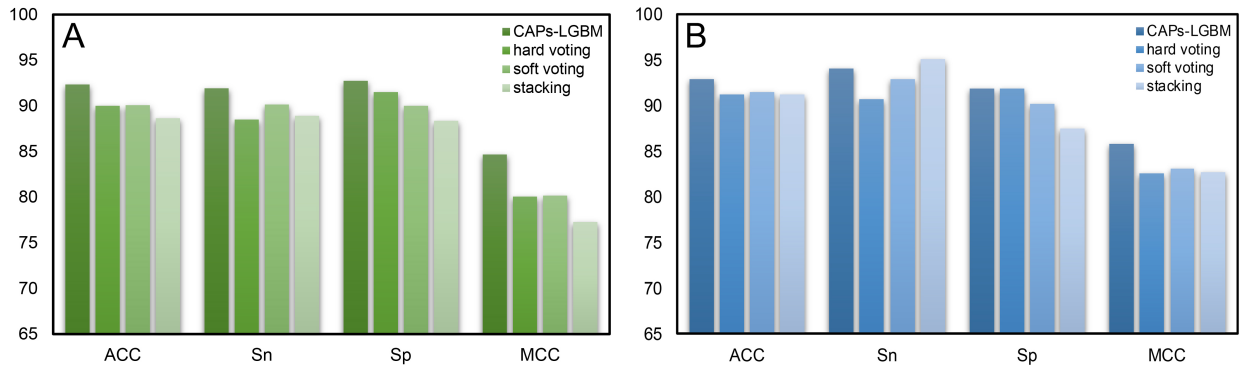

Fig. 7. Comparison results of CAPs-LGBM with three traditional ensemble methods. (A) Training dataset. (B) Test dataset.

sequences in the input box. Finally, new query protein sequences were allowed to enter the input box. In addition, detailed instructions and example sequences in fasta format can be found on the web server interface. The home page provides links of the contact information of authors and relevant data to download.

4. Conclusions

Channel proteins are proteins that can transport molecules past the plasma membrane through free diffusion movement. They can be divided into ion channel proteins and aquaporins. Because it is generally located in the membrane system, the protein structure usually contains an α-helix. In this study, the samples were subjected to enrichment and network analysis. The subnetwork detected by the MCODE tool usually consisted of several single channel proteins. For example, MCODE 1 is a subnetwork construct with several potassium channel proteins, which are tetrameric ion channels existing in all cell types. Potassium channels control the resting membrane potential of neurons, help to regulate the action potential of the myocardium and release insulin by pancreatic cells. Cluster results suggested that passive transport by aquaporins (R-HSA-432047) is significant in humans and mice. Aquaporin is a six-channel transmembrane protein that forms channels on the cell membrane. Each monomer contains a central channel composed of two asparagine-proline-alanine motifs (NPA boxes), which determine water and/or solute selectivity. AQP0/MIP, AQP1/2/3/4/5/7/8/9, and AQP10 transport water inside and outside the cell through the osmotic gradient of the cell membrane. Four aquaporins (AQP7/9 and AQP10) conduct urea, four aquaporins (AQP3/7/9 and AQP10) conduct glycerol, and aquaporin AQP6 conducts anions, especially nitrates. AQP8 conducts water and ammonia. Knockout mice lacking AQP11 formed fatal cysts in the proximal renal tubules. Exogenous AQP12 was localized intracellularly and expressed only in pancreatic acinar cells. Therefore, the biological function of channel proteins is very important. Developing a tool to identify channel proteins is necessary for biological research on channel proteins.

We screened reliable and experimentally verified channel protein sequences as the dataset and established an effective prediction classifier on this basis. First, we selected 17 feature coding methods and four machine learning classifiers to generate 68-dimensional data probability features. Then, the two-step feature selection strategy was used to optimize the features, and the final prediction Model M16-LGBM was obtained on the 16-dimensional optimal feature vector. We made a comprehensive comparison between the proposed probability feature and the existing sequence-based feature description. The results showed that the probability features generated by the feature representation learning method had strong resolution and that the CAPs and non-CAPs were easier to separate. In addition, the proposed CAPs-LGBM was compared with three common ensemble learning strategies to verify the feature representation learning scheme. The 10-fold CV and the independent test showed that CAPs-LGBM was superior to other methods in CAP prediction. Finally, we also established a user-friendly online predictor to promote the use of relevant research communities. We expect that CAP-LGBM will contribute to the identification of CAPs, reveal their biological function mechanism, and accelerate pathological research related to CAPs.

Abbreviations

CAPs, channel proteins; non-CAPs, non-channel proteins; RF, random forest; XGBT, XGBoost; LGBM, light gradient boosting machine; SVM, support vector machines; PPI network, protein-protein interaction network; CLIC1, chloride intracellular channel 1; ITGαV, integrin alpha v; ITGβ1, integrin Beta 1; AQP4, Aquaporin-4; HERG, human ether - à - go related gene; SCLC, small cell lung cancer; VGSCs, voltage-gated Na^+ channels; PCA, prostate cancer; CD-Hit, Cluster Database at High Identity with Tolerance; Acc, accuracy; SE, sensitivity; SP, specificity; MCC, Matthew correlation coefficient; ROC, Receiver operating characteristic; AUC, area under the ROC curve; MF, molecular function; BP, biological process; CC, cell composition; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; AAC, amino acid compo-

sition; DPC, dipeptide composition; PAAC, pseudo-amino acid composition; APAAC, amphiphilic pseudo-amino acid composition; DDE, dipeptide deviation from expected mean; ASDC, adaptive skip dinucleotide composition; CKSAAGP, composition of k-spaced amino acid group pairs; CTD, composition (C)-transition (T)-distribution (D) model; GAAPC, grouped amino acid peptide composition; CTriad, conjoint triad; QSOrder, quasi-sequence order.

Author Contributions

LX, QZ and LZ designed the research; ZC, AELH and SJ performed the research; ZC, MS and DZ analyzed the data; ZC wrote the manuscript. All authors read and approved the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

The work was supported by the National Natural Science Foundation of China (No.62002244), Research foundation of Shenzhen Polytechnic (6021310016K) and the Post-doctoral Foundation Project of Shenzhen Polytechnic China (No. 6020330002K, 6022310029K, 6020330001K, 6021330003K).

Conflict of Interest

The authors declare no conflict of interest. QZ is serving as the guest editor and the editorial board member of this journal. We declare that QZ had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to GP.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2706177.

References

[1] Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International Journal of Cancer. 2010; 127: 2893–2917.

[2] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: a Cancer Journal for Clinicians. 2011; 61: 69–90.

[3] Zhuang J, Dai S, Zhang L, Gao P, Han Y, Tian G, *et al*. Identifying Breast Cancer-induced Gene Perturbations and its Application in Guiding Drug Repurposing. Current Bioinformatics. 2020; 15: 1075–1089.

[4] Feng J, Xu J, Xu Y, Xiong J, Xiao T, Jiang C, *et al*. CLIC1 promotes the progression of oral squamous cell carcinoma via integrins/ERK pathways. American Journal of Translational Research. 2019; 11: 557–571.

[5] Simone L, Pisani F, Mola MG, De Bellis M, Merla G, Micale L, *et al*. AQP4 Aggregation State is a Determinant for Glioma Cell Fate. Cancer Research. 2019; 79: 2182–2194.

[6] Glassmeier G, Hempel K, Wulfsen I, Bauer CK, Schumacher U, Schwarz JR. Inhibition of HERG1 K+ channel protein expression decreases cell proliferation of human small cell lung cancer cells. PflüGers Archiv - European Journal of Physiology. 2012; 463: 365–376.

[7] Bugan I, Kucuk S, Karagoz Z, Fraser SP, Kaya H, Dodson A, *et al*. Anti-metastatic effect of ranolazine in an in vivo rat model of prostate cancer, and expression of voltage-gated sodium channel protein in human prostate. Prostate Cancer and Prostatic Diseases. 2019; 22: 569–579.

[8] Lv H, Dao FY, Zulfiqar H, Su W, Ding H, Liu L, *et al*. A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. Briefings in Bioinformatics. 2021; 22: bbab031.

[9] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics. 2018; 34: 4007–4016.

[10] Song B, Li F, Liu Y, Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Briefings in Bioinformatics. 2021; 22: bbab282.

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

[12] Tang Y, Pang Y, Liu B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics. 2021; 36: 5177–5186.

[13] Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: a Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. Molecular Therapy - Nucleic Acids. 2019; 16: 733–744.

[14] Rostami M, Forouzandeh S, Berahmand K, Soltani M, Shahsavari M, Oussalah M. Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artificial Intelligence in Medicine. 2022; 123: 102228.

[15] Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Medicinal Research Reviews. 2020; 40: 1276–1314.

[16] Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. Briefings in Functional Genomics. 2021; 20: 1–18.

[17] Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics. 2020; 88: 397–413.

[18] Zhang D, Chen H, Zulfiqar H, Yuan S, Huang Q, Zhang Z, *et al*. IBLP: an XGBoost-Based Predictor for Identifying Bioluminescent Proteins. Computational and Mathematical Methods in Medicine. 2021; 2021: 6664362.

[19] Liu B, Gao X, Zhang H. 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. 2019; 47: e127.

[20] Shao J, Yan K, Liu B. FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. Briefings in Bioinformatics. 2021; 22: bbaa144.

[21] Wang X, Gao P, Liu Y, Li H, Lu F. Predicting Thermophilic Proteins by Machine Learning. Current Bioinformatics. 2020; 15: 493–502.

[22] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, *et al*. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021; 49: D480–D489.
[23] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, *et al*. The Pfam protein families database. Nucleic Acids Research. 2012; 40: D290–D301.
[24] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152.
[25] Zou Q, Lin G, Jiang X, Liu X, Zeng X. Sequence clustering in bioinformatics: an empirical study. Briefings in Bioinformatics. 2018. (in press)
[26] Chen Y, Ma T, Yang X, Wang J, Song B, Zeng X. MUFFIN: multi-scale feature fusion for drug-drug interaction prediction. Bioinformatics. 2021; 37: 2651–2658.
[27] Yang W, Zhu X, Huang J, Ding H, Lin H. A Brief Survey of Machine Learning Methods in Protein Sub-Golgi Localization. Current Bioinformatics. 2019; 14: 234–240.
[28] Shao J, Liu B. ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm. Briefings in Bioinformatics. 2021; 22: bbaa192.
[29] Liu B, Zhu Y, Yan K. Fold-LTR-TCP: protein fold recognition based on triadic closure principle. Briefings in Bioinformatics. 2020; 21: 2185–2193.
[30] Wei H, Xu Y, Liu B. ICircDA-LTR: identification of circRNA-disease associations based on Learning to Rank. Bioinformatics. 2021; 37: 3302–3310.
[31] Zhang YP, Zou Q. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics. 2020; 36: 3982–3987.
[32] Zhang G, Yu P, Wang J, Yan C. Feature Selection Algorithm for High-dimensional Biomedical Data Using Information Gain and Improved Chemical Reaction Optimization. Current Bioinformatics. 2020; 15: 912–926.
[33] Rao B, Zhou C, Zhang G, Su R, Wei L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Briefings in Bioinformatics. 2020; 21: 1846–1855.
[34] Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics. 2020; 36: 3350–3356.
[35] Zhang J, Zhang Z, Pu L, Tang J, Guo F. AIEpred: an Ensemble Predictive Model of Classifier Chain to Identify Anti-Inflammatory Peptides. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021; 18: 1831–1840.
[36] Hu Y, Zhang H, Liu B, Gao S, Wang T, Han Z. rs34331204 regulates TSPAN13 expression and contributes to Alzheimer's disease with sex differences. Brain. 2020; 143: e95.
[37] Hu Y, Sun J, Zhang Y, Zhang H, Gao S, Wang T, *et al*. rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. BMC Medicine. 2021; 19: 11.
[38] Hu Y, Qiu S, Cheng L. Integration of Multiple-Omics Data to Analyze the Population-Specific Differences for Coronary Artery Disease. Computational and Mathematical Methods in Medicine. 2021; 2021: 7036592.
[39] Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, *et al*. ILearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in Bioinformatics. 2020; 21: 1047–1057.
[40] Zeng X, Zhong Y, Lin W, Zou Q. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Briefings in Bioinformatics. 2020; 21: 1425–1436.
[41] Zeng X, Zhu S, Hou Y, Zhang P, Li L, Li J, *et al*. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. Bioinformatics. 2020; 36: 2805–2812.
[42] Zhang D, Xu Z, Su W, Yang Y, Lv H, Yang H, *et al*. ICarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics. 2021; 37: 171–177.
[43] Jin Q, Cui H, Sun C, Meng Z, Su R. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. Knowledge-Based Systems. 2021; 218: 106753.
[44] Liu J, Su R, Zhang J, Wei L. Classification and gene selection of triple-negative breast cancer subtype embedding gene connectivity matrix in deep neural network. Briefings in Bioinformatics. 2021; 22: bbaa395.
[45] Wei L, Luan S, Nagai LAE, Su R, Zou Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics. 2019; 35: 1326–1333.
[46] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics. 2002; 35: 352–359.
[47] Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and Structural Biotechnology Journal. 2020; 18: 153–161.
[48] Su R, Wu H, Xu B, Liu X, Wei L. Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019; 16: 1231–1239.
[49] Dao F, Lv H, Wang F, Feng C, Ding H, Chen W, *et al*. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. 2019; 35: 2075–2083.
[50] Zhang J, Xiong Y, Min S. A new hybrid filter/wrapper algorithm for feature selection in classification. Analytica Chimica Acta. 2019; 1080: 43–54.
[51] He S, Guo F, Zou Q. MRMD2.0: A Python tool for machine learning features ranking and reduction. Current Bioinformatics. 2020; 15: 1213–1221.
[52] Hong Z, Zeng X, Wei L, Liu X. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics. 2020; 36: 1037–1043.
[53] Manavalan B, Basith S, Shin TH, Wei L, Lee G. MAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. 2019; 35: 2757–2765.
[54] Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q. Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2014; 11: 192–201.
[55] Wang H, Tang J, Ding Y, Guo F. Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. Briefings in Bioinformatics. 2021; 22: bbaa409.
[56] Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. International Journal of Data Mining and Bioinformatics. 2013; 8: 282–293.
[57] Huang Y, Zhou D, Wang Y, Zhang X, Su M, Wang C, *et al*. Prediction of transcription factors binding events based on epigenetic modifications in different human cells. Epigenomics. 2020; 12: 1443–1456.
[58] Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylo-

genetic estimates. Nature Methods. 2017; 14: 587–589.
[59] Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010; 59: 307–321.
[60] Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018; 35: 518–522.
[61] Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution. 2013; 30: 1188–1195.
[62] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications. 2019; 10: 1523.
[63] van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008; 9: 2579–2605.