International Journal of Nanotechnology

ISSN online: 1741-8151 - ISSN print: 1475-7435
https://www.inderscience.com/ijnt

Predictive protein module based on PPI network and double clustering algorithm

Sicong Huo, Quansheng Liu, Tao Lu

DOI: 10.1504/IJNT.2024.10062015

Article History:

Received:	10 March 2021
Last revised:	15 April 2021
Accepted:	30 June 2021
Published online:	05 February 2024

Predictive protein module based on PPI network and double clustering algorithm

Sicong Huo, Quansheng Liu and Tao Lu*

College of information engineering,
Nanning University,
Nanning, Guangxi, 530200, China
Email: huike781457huanl@163.com
Email: duanchi5027fudu@163.com
Email: lutao_237@outlook.com
*Corresponding author

Abstract: Protein-protein interaction (PPI) is a kind of biomolecular network which plays an important role in biological activities. In order to improve the accuracy of protein function module prediction, obtain protein function module and run timely, this paper proposes predictive protein module based on PPI network and double clustering algorithm (ISCC), which considers the characteristics of PPI network and considers nodes as two-dimensional data points. First of all, improved density-based spatial clustering of applications with noise (IDBSCAN) determines the central cluster, and then uses spectral clustering (SC) to redivide the weight; secondly, CFSFDP and Chameleon algorithm are used to filter the similarity of the central cluster, and use support vector machine (SVM) to get the final clustering result. Finally, the experiments are compared with CDUN, EA, MCL and MCODE in terms of accuracy, sensitivity, *F* value and the number of protein functional modules. The experimental results show that the *F* value of ISCCD is 70% higher than that of EA, the number of recognition modules is 257 higher than that of CDUN, and the running time is 494s faster than that of MCODE.

Keywords: PPI network; protein function module prediction; clustering; SVM; support vector machine; IDBSCAN; improved density-based spatial clustering of applications with noise.

Reference to this paper should be made as follows: Huo, S., Liu, Q. and Lu, T. (2024) 'Predictive protein module based on PPI network and double clustering algorithm', *Int. J. Nanotechnol.*, Vol. 21, Nos. 1/2, pp.112–128.

Biographical notes: Sicong Huo is currently pursuing a PhD in Information Technology at Segi University and working at Nanning University. His primary research focus is the prediction of protein functional modules in bioinformatics, specifically in the context of protein-protein interaction (PPI) networks, and how to accurately predict key protein functional modules through dynamic PPI networks.

Quansheng Liu, PhD, is working at Nanning University. He is highly interested in data mining and currently specialises in predicting functional features through data mining.

Tao Lu is currently pursuing a PhD in Educational Technology Innovation at KMITL University. He works at Nanning University and is interested in computer algorithm optimisation. He is currently focusing on using optimisation algorithms to predict protein functional modules.

1 Introduction

Human life is inseparable from protein, which is also an important part of the organism. There are many kinds and different functions of protein and almost all life activities need to be completed through protein [1]. It is a dynamic process for protein's living space and time, while in this process, there will be interaction between proteins. This interaction forms a protein-protein interaction (PPI) network [2]. In a PPI network, protein functional modules are a collection of proteins that interact in different time and space stages to complete a specific molecular process [3]. Protein function module prediction is the main trend in the field of modern biological computing science. A large number of biological computing experiments produce a large number of protein-protein interaction data, which is the basis of protein function module prediction. Protein function module helps to understand the protein function, tissue structure, physiological ability and function of cells, etc. Therefore, it is of great significance to predict protein functional modules [4].

At present, various optimisation algorithms have been used to predict protein functional modules in the study of protein functional modules. There are many framework applications of machine learning to predict protein function modules in medical and biological computing fields [5–7]. According to the calculation characteristics of the algorithm for predicting protein function modules, the algorithms for predicting protein function modules are divided into density-based clustering algorithm, hierarchical clustering algorithm [8,9], partition-based clustering algorithm [10,11], intelligent population based and PPI network clustering algorithm [12,13]. However, hierarchical clustering algorithm is difficult to detect the overlapping protein functional modules in PPI network, and the clustering results are greatly affected by noise; the algorithm based on partition clustering needs to determine the number of clusters manually, which cannot predict the protein functional modules well, and a large number of false positives will appear in the clustering results; there is a contradiction between intelligent population and PPI network clustering algorithm. The local search ability of PPI network is poor, and it is easy to produce data noise; while the density based clustering prediction algorithm, which is not limited to the distribution shape of the original data, can converge to the global optimal solution. Therefore, the density-based algorithm has been successfully applied to the PPI network function module mining, and has become a research hot spot in this field. Lashkov et al. [14] proposed the application of the DBSCAN algorithm to detect hydrophilic clusters in protein structures, which provides an algorithm framework for predicting protein function. Aiming at the problem of using DBSCAN algorithm to predict protein functional module clustering, the related verification was carried out. Melvin et al. [15] group proposed visualising correlated motion with HDBSCAN clustering, although the method it propose here can be used to group residues using correlation as a similarity matrix (the most straightforward and

intuitive method) it can also be used to more generally determine groups of residues which have similar dynamic properties. As they are based not on spatial closeness but rather closeness in the column space of a correlation matrix. Iwendi et al. [16] proposed that the purpose of the improved KMT with ACO is to design an appropriate route to connect the starting point and ending point of the environment with intruders, it can effectively balance the convergence speed and communication of the sensor nodes. In order to improve the search path, Mittal et al. [17] proposed a LEACH protocol with Levenberg-Marquardt neural network (LEACH-LMNN) is considered to analyse the overall network lifetime [17]. In order to improve the quality of predicted protein function modules, the use of ensemble models for multiple class and binary class classification for improving introduction detection systems, which improves the classification efficiency [18].

Although the mining of functional modules based on density clustering algorithm and PPI network has achieved certain results, how to effectively construct uncertain PPI, how to overcome the defects of low accuracy, low sensitivity, low efficiency and so on caused by the sensitivity of clustering centre and clustering number, are still urgent problems to be solved. The main contribution of this paper is to propose predictive protein module based on PPI network and double clustering algorithm (ISCC) to improve the quality of clustering centres and the number of clusters, to improve the accuracy and sensitivity of predicting protein function modules, and to obtain the number of protein function modules, so as to improve the experimental running time, S precision (SP), recall sensitivity (Sn) and *F*-measure in the number of experiments. The organisation of this paper is as follows: firstly, the PPI Network is constructed from a new structure, and the degree of each node in the PPI network is divided again. Secondly, the improved density based spatial clustering of applications with noise is used to select the centre cluster. Thirdly, the special clustering is used to optimise the centre cluster. Finally, the SVM classifier is used to classify and get the clustering results. The experimental results show that the proposed predictive protein module based on PPI network and double clustering algorithm (ISCCD) can improve the *F* value by 70% compared with EA, the number of recognition modules are 257 higher than CDUN and the running time is 494 s faster than MCODE.

2 Problem and definition

Definition 1: PPI network is usually expressed as an undirected graph $G(V, E)$ [19], where $V = \{v_1, v_2, v_3, \ldots, v_n\}$, expressed as a set of nodes, $E = \{e_1, e_2, e_3, \ldots, e_m\}$ is expressed as the set of connected edges of each node. Node v_i (1, 2, 3, …, m) is represented as a PPI network node, and e_m is represented as the interaction between PPI network nodes and nodes. In this paper, the network structure, the location of each node, and the core node in PPI are regarded as important factors for predicting protein function modules. The matrix $V_{n\times m}$ is represented as a graph G, where each node v_{ij} (i = 1, 2, 3, 4… …N, j = 1, 2, 3, …, m) is defined as follows:

$$v_{ij} = \begin{cases} 1, e_{ij} \in E \\ 0, e_{ij} \notin E \end{cases} \tag{1}$$

Definition 2 The matrix $Y_{n\times z}$ represents the annotation information label of the protein database that has been successfully identified, where z is the type of known protein functional modules, and n is the label to be compared. It is defined as follows:

$$Y_{ij} = \begin{cases} 1, \\ 0 \end{cases} \tag{2}$$

According to the above definition, the problem of predicting protein function module is transformed into a multi-label two classification problem, and the mapping function H: $X \rightarrow Y$ is obtained, thereby obtaining the predictive protein function module.

3 ISCC algorithm

In order to improve the prediction accuracy of protein energy supply module, the prediction algorithm has some disadvantages, such as low accuracy, big influence of data noise and small number of protein function modules. This paper proposes predictive protein module based on PPI network and double clustering Algorithm: (1) using dip database to download PPI network data, using the degree of PPI network nodes to reconstruct PPI network, and then using software to divide the new PPI network graph; (2) considering the characteristics of PPI network misdirected graph, the nodes are regarded as two-dimensional data points, and using the improved density based spatial clustering algorithm proposed in this paper for the original PPI network nodes clustering of applications with noise, (3) according to the spectral clustering (SC) algorithm proposed in this paper, the weight of the determined cluster is divided again to optimise the central cluster, reduce the situation of falling into local optimum, and improve the accuracy of protein function module prediction; (4) CFSFDP [20] and Chameleon [21] algorithms are used to filter the similarity of central clusters and get the final cluster. Finally, SVM classifier is used for classification.

3.1 Structure of the PPI network

Due to the limitations of experimental conditions and the characteristics of PPI network topology, PPI network has the influence of uncertainty in predicting protein functional modules, and it is easy to appear false-positive phenomenon in the experimental results. [22] if it directly uses the original PPI network data, it will be greatly affected by the noise, and it needs to process the PPI network data to improve the accuracy of protein function module prediction. Firstly, the PPI network is downloaded from DIP [23] database, and then the degree (ECC) of interaction between two nodes in PPI network is calculated by formula (3), where H_c is the number of all neighbour nodes of node u and v, d_u and d_v are the degree of node u and v. Using the value of degree to measure the interaction of each group, a new PPI network diagram is formed, and then the PPI network is drawn by the software Cytoscape [24]. According to Figure 1, the PPI network downloaded from dip dataset is described. This network only shows the interaction of 10 proteins and 16 proteins, and each line represents their interaction. Figure 2 is a new PPI network structure based on the measurement of each interaction, which is calculated by the edge aggregation coefficient. Each interaction is re divided by the degree of the edge. The value of each edge represents their compactness. The smaller the number of

two nodes is, the greater the interaction between nodes is. The larger the interaction, the easier it is to predict the protein function module.

$$\mathrm{ECC} = \frac{H_c}{\min\left(d_u - 1, d_v - 1\right)} \tag{3}$$

Figure 1 PPI network

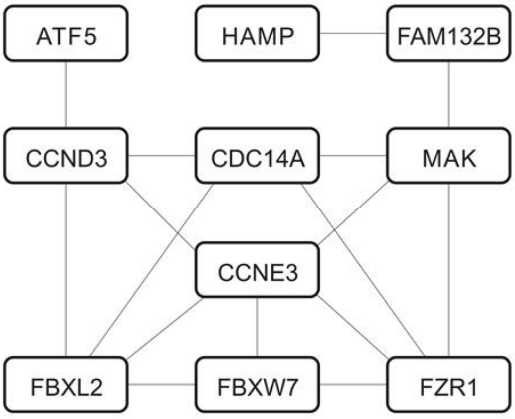
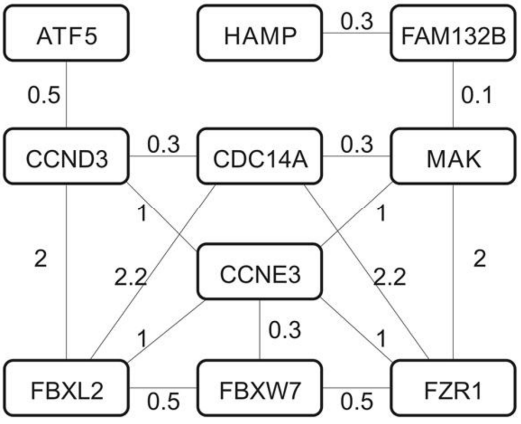


Figure 2 Newly constructed PPI network



3.2 ISCC algorithm analysis

According to the new planning PPI network topology, if the traditional density clustering algorithm is used to predict the protein function module, it is necessary to manually determine the setting of the PPI network clustering centre. The whole clustering process needs to be continuously updated. The division of cluster centres. If there is a deviation in the initial selection of the clustering centre, false positives may occur, resulting in poor clustering results and inconsistent problems with the actual situation, resulting in the process of predicting protein function modules easily falling into a local optimal situation, and finally making the experimental data. The precision and accuracy rate are reduced. To solve this problem, this paper proposes that the improved density clustering

algorithm does not need to manually divide the central node and the maximum number of clusters. The algorithm first calculates the relative distance and connection coefficient between nodes in the newly formed PPI network. Re-divide the weights to obtain the corresponding cluster centres.

After using the improved density clustering algorithm to obtain the central cluster, if the data is directly classified, the noise of the data will be very high and the accuracy is not high. In order to solve these problems, this paper uses the improved density clustering algorithm to obtain a new central cluster. Then use fuzzy spectrum clustering algorithm for clustering analysis, which can improve the number and accuracy of protein function module prediction. This algorithm compresses the repeated redundant rules, and after initially forming the rule population, it regards the rules in the rule population as two-dimensional data points again, and replaces the clustering of the original data points with the clustering of the rule points. One rule can be Represents a large number of data points, which can greatly reduce the amount of calculation required for clustering, and there is the possibility of reducing high-dimensional data to low-dimensional later. For each regular centre cluster, this method reduces the number of iterations of the algorithm to improve the efficiency of the algorithm. At the same time, it can also mine more predictive protein function modules and improve the accuracy of the algorithm. The following is the ISCC algorithm idea as follows:

Step 1: Use formula (3) and (4) to calculate the distance between nodes in PPI network and the density of newly constructed PPI network, calculate the local density ρ_i of v_i and the distance δ_i between v_i and higher density points, d_{ij} and d_c represent the positions of two nodes, and their definitions are as follows:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{4}$$

$$\delta_i = \min_{j:p_i > p_j} (d_{ij} - d_c) \tag{5}$$

Let $G = (v, e, p)$ be an indeterminate graph, where V represents the nodes in PPI network, e represents the edges that interact with each other in PPI network, and p represents the mapping weight of the edge connection distance between any two nodes. In PPI network, $V = \{v_1, v_2, v_3, \ldots, v_n\}$, $E = (e_1, e_2, e_3, \ldots, e_n)$ is represented by nodes, e_u and e_v represent the coefficients of edge degree in clustering. In formula (5), EC (edge coefficient, EC) is the coefficient of edges in PPI network, SC is the number of nodes v and u in PPI network, and v_i and v_j have the distance between nodes. S_V and S_u denote the node distances v_i and v_j of any two points, and there is a sequence between nodes $r = (s_0, s_1, s_2, \ldots, s_n)$, where $S_v \in V(0 \le q \le N)$, $S_U \in V(0 \le q \le N)$. Actual distance (AD) in formula (6) is the actual distance between two nodes, s_v and s_u are the node distances of any two points, and v_i and v_j exist the sequence between nodes $r = (s_0, s_1, s_2, \ldots, s_n)$, where $S_v \in V(0 \le q \le N)$, $S_v \in V(0 \le q \le N)$, then the similarity weight between any two nodes v_i and v_j in PPI network is calculated by formula (7) WS (weight similarity, WS).

$$\mathrm{EC} = \frac{S_c}{\min(v_i, v_j)} \tag{6}$$

$$AD(S_V, S_U) = \frac{1}{\min\limits_{s \in R_{oj}} \sum_1^{|N|-1} (e^{p^2(s_v, s_u)} - 1) + 1} \tag{7}$$

$$\mathrm{WS} = \left(x_i x_j\right) = EC\left(x_{i,} x_j\right) \times AD\left(x_i, x_j\right) \tag{8}$$

Step 2: Based on step 1, the weight density function D_i of the sample is calculated by formula (8), and the weight division at node v_i and v_j is defined as:

$$D_i = \sum_{j=1}^{N} \frac{1}{1 + \dfrac{4WS(v_i, v_j)}{r_j^{\,2}}} \tag{9}$$

Among them, d_r in formula (9) is the neighbourhood radius, and its value is the average distance measure of N objects. Obviously, the closer v_i is to other nodes, the smaller d_r is, and the larger the corresponding weight value is.

$$d_r = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} WS(v_i, v_j)}{N(N-1)}} \tag{10}$$

Step 3: get the initial cluster centre $D_i(k)$ through formula (10), set it as the k cluster centre point ck, where $D_i = \max\left\{D_I^{k-1}, i = 1, 2, 3 \ldots . N\right\}$, then get the cluster centre function as follows:

Step 4: a new distance centre $J(u, c)$ is obtained by formula (11), where de (v_i, v_j are the original cluster centres)

$$J(u, c) = \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ij}^{\ 2} \, DE(v_i, v_j) \tag{11}$$

According to formula (12), u_{ij} is the membership degree of node x_{ij}. According to the iterative optimisation objective function of membership degree $\mathrm{v_i}$ and clustering centre c_j, the minimum objective function c_j is obtained, as shown in formula (13), where v_i is the node and $m > 1$ is the weighted index.

$$u_{ij} = \left[\sum_1^N \left(\frac{de(x_i, c_j)}{de(x_i, c_k)} \right)^{\frac{1}{m-1}} \right]^{-1} \quad i = 1.2.3 \ldots . N; j = 1, 2, 3 \ldots . K; \tag{12}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{\ m} \times v_i}{\sum_{i=1}^{N} u_{ij}^{\ m}}, j = 1.2.3 \ldots . . K \tag{13}$$

3.3 ISCC algorithm aggregation

It can be seen from formula (10) that the closer to the cluster centre point, the smaller the corresponding P value. When D iteration stops, K data points with higher global density can be obtained as the initial cluster centres of the cluster. In this process, the high-density samples rather than the edge outliers are at the centre of the category, so that the

selected centre points of the category belong to different categories as much as possible, and K initial category centre points can be obtained to reduce the influence of noise points on the experimental results. After calculating the distance c_i of each PPI network node c_i, c_i and u_i are artificially selected by the decision graph method. A major problem with this method is that the clustering effect obtained when there are many density extreme points in the data distribution is not ideal. According to the ISCC algorithm objective function, iteratively update the cluster centres and the degree of membership, and optimise the ISCC algorithm sensitive to the initial cluster centres. This paper adopts the method proposed by clustering by fast search and find of density peaks (CFSFDP) [20], and further optimises the points where the cluster centre selection square value is higher. The cluster centre can be gradually calculated as shown in formulas (14) and (15). FD_i represents the expected cluster centre, TD_i is the true cluster centre, $\sigma(c_j)$ is the standard deviation of the distance calculated by equation (13), and u_i is the average membership degree of v_i. According to the CFSFDP algorithm, the distance between each cluster centre is relatively long, so the cluster centre data distance between points should satisfy equation (14). At the same time, since the CFSFDP algorithm will produce a noise cluster centre c_j with a larger distance u_i and a smaller density, it is necessary to use equation (15) to separate the noise of the expected cluster centre. When the u_i and c_i of the rule v_i satisfy the above formulas (14) and (15), the point can be considered as the centre of the cluster.

$$FD_i = (u_i) \geq 2\sigma(c_i) \tag{14}$$

$$TD_i = FD_i \geq \mu(c_i) \tag{15}$$

Since the number of initial sub-categories generated by satisfying FD_i and TD_i conditions is large, the algorithm in this paper uses the Chameleon algorithm [21] to calculate the distance between the initial sub-categories according to formula (16) and (17), which is mainly for gathering in PPI networks. The central cluster formed after the class is screened. The main criterion for screening is to measure and compare the similarity between classes, because the performance between nodes and the data are normalised during data input to cluster the similarities. The sub-categories are merged to obtain the final clustering result. In the process of fusing the sub-categories, the Chameleon algorithm not only considers the relative closeness between the two clusters RI_{ij} (the relative interconnection between cluster i and cluster j), but also considers two internal connection of the cluster RC_{ij} (the relative similarity between cluster i and cluster j). Among them, AI_{ij} (cluster link after PPI clustering) absolute interconnectivity is the sum of the weights of edges connecting cluster G_i node i to cluster G_j node. AI_i or AI_j (absolute internal interconnection) is the sum of the weights of the edges belonging to the smallest dicing bisector of cluster G_i or cluster G_jAC_{ij} (absolute approximation) is the average of the edges connecting the points in cluster G_i to the points in cluster G_j. The weight AC_i or AC_i (absolute internal proximity) is the average weight of the edges belonging to the smallest split bisector of cluster G_i or cluster G_j. After the formula (TD), the core cluster of the protein function module is finally predicted.

$$RI_{\mathrm{ij}}=AI_{\mathrm{ij}} / [AI_i + AI_j) / 2] \tag{16}$$

$$RC_{ij} = \frac{AC_{ij}}{\frac{|G_i|}{|G_i + G_j|AC_i} + \frac{|G_j|}{|G_i + G_j|AC_i}} \tag{17}$$

$$TD_{ij} = \sum_{i,j=1}^{N} (RC_{ij} * RI_{ij}) \tag{18}$$

3.4 ISCC algorithm classification

Protein function prediction can be regarded as a multi-label binary classification problem. A protein (sample) may have multiple functions (labels) that have a semantic relationship with each other at the same time. Considering the correlation between tags, this paper uses SVM as the classifier [8]. Compared with traditional classifiers, the use of SVM is more flexible. It can be modelled by setting the number of nodes, activation function and loss function of the output layer, and the hidden layer is used to mine the complex relationship between tags. In this paper, the known protein functional module terminology and the cluster data standardised by ISCC algorithm are used as feature vectors as input, and the functional annotation vectors are used as output. The linear mapping capability of SVM is used to mine the correlation between functional annotation terms and establish multiple features. The mapping relationship between multi-function and protein function module prediction.

The specific content of the ISCC algorithm is as follows:

Algorithm process 1:

Input: PPI network cluster $G(V, E)$ dataset, with identified protein functional module terms.

Output: predict protein function module Y

Algorithm process 2: Use the improved density clustering algorithm to perform cluster analysis on the PPI network, and appropriately compress the nodes

2.1) Use the improved density clustering algorithm proposed in this paper to calculate the local density and edge coefficients of each PPI network cluster

2.2) Determine the centre of the cluster

2.3) Divide each cluster into the central cluster with the shortest cluster and the highest local density.

Algorithm process 3: re partition the weight of the centre cluster with the highest density

3.1) the spectral clustering algorithm is proposed to calculate the local density and edge coefficient of each PPI network cluster

3.2) determine the centre of the cluster.

Algorithm process 4: Properly aggregate the planned ethnic groups.

4.1) The clusters generated in step 3.2 are calculated using formulas (14)–(18) TD

4.2) Plan the TD result with the highest value and merge each cluster

4.3) Repeat steps 4.1 and 4.2 until the termination condition is reached.

4.4) The core cluster of predicted protein function

Algorithm process 5: Perform SVM classification

5.1) Classify the core clusters of predicted protein function modules obtained by algorithm process 4 and known protein terms

5.2) Predict protein function module to get function Y

4 Experimental results and analysis

4.1 Experimental data and evaluation indicators

The environment of this experiment is Windows 10 operating system, I7 data processor, and the programming tool is pychram. In order to verify the effectiveness of ISCC algorithm proposed in this paper, yeast PPI interaction network data in database of interacting protein (DIP) [24] was selected as experimental data. The database contains 5002 proteins and 21,559 pairs of interaction data (DIP) It is convenient to use keywords such as gene name. The results of the query list two items: node and link. The node describes the characteristics of the protein, including the domain and fingerprint of the protein. The PPI network of *Saccharomyces* is relatively stable. Go functional protein annotation provides gene ontology (GO) [9] which can make the functional description of gene products in dip database consistent. The definition of GO has been used in many protein databases, which makes the query in these databases highly consistent. This kind of definition language has multiple structures, so it can be used to query in various degrees. For example, GO can be used to search for gene products related to signal transduction in yeast genome, and also to find various biological receptor tyrosine kinases. At the same time, GO database also provides biological process (BP) and molecular function (MO), The three independent sub ontologies of BP and MF, cellular component (CC), respectively describe the molecular functions that gene products may perform, the cellular environment they live in, and the biological processes they participate in, so that BP and MF are independent of each other. Then the PPI interaction structure network of these two BP and MF is constructed.

In this paper, Sprecision (SP), recallSensitivity (Sn), and *F*-measure are used to measure the prediction effect of the algorithm [10]. Sp refers to the proportion of successfully matched modules among the function modules identified by the algorithm in the number of mined modules, such as formula (19). TP refers to the number of predicted protein modules matched with known protein function modules, and FP refers to the number of predicted protein modules matched with known protein function modules. It represents the number of unmatched functional modules in the predicted protein, S_n represents the proportion of successfully matched functional modules in the benchmark module, as shown in formula (20), and FN represents the number of unmatched functional modules in the predicted protein module and the benchmark module. F-measure is used to evaluate the overall performance of the algorithm. In order to avoid the bias caused by sensitivity and specificity, the formula (21) is as follows:

$$Sp = \frac{TP}{TP + FP} \tag{19}$$

$$Sn = \frac{TP}{TP + FN} \tag{20}$$

$$F\text{-Measure} = \frac{2 * Sp * Sn}{Sp + Sn} \tag{21}$$

For the selection of protein function modules, the matching degree between the function module A and the known function module B extracted by the algorithm is calculated by formula (22) to obtain the prediction function module result. If the matching degree between the identified functional module A and the known functional module B exceeds a given threshold, the known functional module is said to be identified, and the threshold is set to 0.2 [2160–2162] according to the document [22]. If $OS(A, B) \geq 0.2$, it is to predict that the protein function module is recognised by the known protein function module.

$$OS(A, B) = \frac{|A \cap B|}{|A||B|}. \tag{22}$$

4.2 Analysis of results

In order to verify the performance of the ISCC algorithm proposed in this paper, the ISCC algorithm and CDUN [25], EA [26], MCL [27], MCODE [28] and other four algorithms are independently run 20 times, and the average of all experimental results is taken. Analyse, get the data of each algorithm for PPI network mining protein function module. This paper firstly presents the characteristics of PPI network function modules of MF, BP, and CC, and then each algorithm mines the basic information of the function modules, then gives the prediction effect of each algorithm, and finally draws a conclusion based on the experimental results.

This paper firstly presents the feature extraction of functional module features and attribute principal components of the three PPI networks of MF, BP, and CC, and the setting of SVM nodes, and then compares and analyses its prediction effects with CIA [29] and GoDIN [30]. As shown in Table 1, the number of protein function modules extracted by the ISCC algorithm on the three PPI networks and the reading of NG modules are not very affected, which proves that the ISCC algorithm will not be affected by the special PPI network and cause the number of acquisition modules to change.

Table 1 Features of functional modules of MF, BP, CC PPI network

GO	*K*	*Modularity*
MF	17	0.825
BP	18	0.832
CC	18	0.831

PM represents the total number of functional modules mined by the algorithm, and FULL refers to the number of fully identified functional modules in a set of known functional modules. It can be seen from Table 2 that 312 of the protein function modules mined by the ISCC algorithm are matched. Among the 5 algorithms, the number of protein function modules is the most mined. Compared with other algorithms, it proves that the protein function module mining algorithm has higher performances efficiency.

Table 2 Basic information of the functional modules of each algorithm mining

Algorithm	*PM*	*FULL*	*TP*
CDUN	123	6	53
MCL	232	10	134
EA	353	8	137
MCODE	257	9	68
ISCC	380	19	260

As shown in Table 3, PM represents the total number of function modules mined by the algorithm, and SC represents the protein function modules with significant significance for each algorithm mining. ISCC algorithm mining significant protein function modules is 312, of which the proportion is 82.11%, compared with other algorithms CDUN, MCL, EA, MCDE 47.15%, 66.81%, 59.77%, 75.09% have improved, which shows that, ISCC algorithm has relatively strong biological computing significance, and it also proves that ISCC algorithm has higher efficiency for mining protein functional module algorithms.

Table 3 Significance statistics of the functional modules of each algorithm mining

Algorithm	*PM*	*SC*	*Proportion (SC/PM)*100*
CDUN	123	58	47.15%
MCL	232	155	66.81%
EA	353	211	59.77%
MCODE	257	193	75.09%
ISCC	380	312	82.11%

In order to further compare the execution efficiency of each algorithm, under the optimised parameters of each algorithm, run 20 times independently in the dataset downloaded by the DIP database, and take the average value of each algorithm to compare the operating efficiency of each algorithm, as shown in Table 3.

As shown in Table 4, when each algorithm mines the module scale range, the number of protein functional modules is greater than 3. The matching rate here refers to the ratio of the number of protein functional modules to the number of standard protein functional modules. The table shows, ICSS algorithm experiment running time is 670.576 s, the running time complexity is the lowest, other algorithms have exceeded the time of 900 s, it can be seen that the algorithm proposed in this paper is more efficient.

Table 4 Comparison of the efficiency of each algorithm

Algorithm	*PM*	*Module scope*	*Matching rate (TP/PM)*	*Run time* (s)
CDUN	123	4-32	42.86%	1537.843
MCL	232	4-143	57.65%	957.247
EA	353	4-132	53.41%	963.125
MCODE	257	4-98	26.27%	1164.135
ISCC	380	4-235	68.47%	670.576

To further prove the efficiency of ISCC algorithm, ISCC algorithm and other algorithms are run 20 times on Krogan [31] dataset, which contains 4562 yeast protein data. After removing self-interaction and repeated interaction, the dataset has 3672 proteins and 14317 proteins. The average running efficiency of each algorithm obtained from Krogan dataset is shown in Table 5. Compared with the DIP data scale, the execution efficiency and matching rate of each algorithm are relatively improved when Krogan dataset with smaller Krogan data scale runs each algorithm. Specifically, the matching rate of CDUN algorithm mining module increased by 1.76%, and the running time decreased by 116.684 s; the matching rate of MCL algorithm mining module increased by 0.062%, and the running time decreased by 360.862 s; the matching rate of EA algorithm mining module increased by 0.070%, and the running time decreased by 91.45 s; the matching rate of MCODE algorithm increased by 1.92%, and the running time decreased by 140.788 s The matching rate of ISCC mining module is improved by 3.89%, and the running time is reduced by 197.919 s. It can be seen from the analysis in Table 5 that the execution efficiency of MCL algorithm and ISCC in processing large datasets is relatively high, while the data size has little effect on the execution efficiency of MCODE and CDUN and EA algorithm. As shown in Tables 4 and 5, the size of data does not affect the matching rate and timeliness of ISCC algorithm, which proves that ISCC algorithm has higher efficiency.

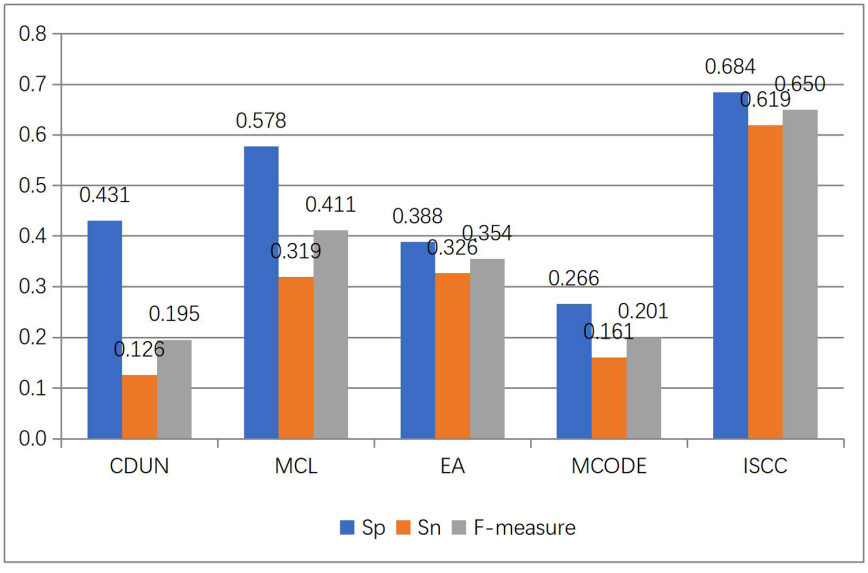
Table 5 Efficiency analysis of each algorithm on Krogan dataset

Algorithm	*PM*	*Module scope*	*Matching rate (TP/PM)*	*Run time* (s)
CDUN	132	4-36	44.62%	1421.159
MCL	272	4-153	63.85%	596.385
EA	383	4-105	54.12%	871.674
MCODE	269	4-92	28.19%	1023.347
ISCC	392	4-248	72.36%	472.657

As shown in Figure 3, the performance of ISCC is compared with the other four algorithms in DIP dataset. The five function prediction methods based on PPI network provided by DIP database are compared in *Sn*, *Sp*, *F*-measure values. The accuracy *Sp* of ISCC algorithm is 0.684, which is 58.7% higher than CDUN, MCL 18.3%, EA 76.2%, MCODE 157.1% higher than CDUN, and the sensitivity is *Sn* compared with CDUN, *F*-measure increased by 391.2%, MCL 94.0%, EA 89.8%, MCODE 284.3% compared with CDUN, it can improve 233.3%, MCL 58.1%, EA 83.6%, MCODE 223.3%.

In general, the ISCC algorithm proposed in this paper can effectively predict the number of protein functional modules, and has high accuracy *Sp*, sensitivity *Sn*, and *F* value compared with other algorithms, which proves that the correct part of protein functional modules that help me to recognise is higher.

Figure 3 Comparison of the performance of protein function prediction algorithms of various algorithms



5 Conclusion

At present, the limitations of PPI network prediction of protein functional modules are: (1) building a new PPI network based on the topological characteristics of PPI network and the biological information of protein has too high dependence on PPI network; (2) there is no good method to integrate biological data to predict the number of protein functional modules. (3) SVM classifier is a binary classifier. The following aspects can be used for in-depth research on future research: (1) Performance analysis of compression technologies for chronic sound image transmission under smart phone enabled tele-sound network can effectively reduce data noise [32], (2) An AI based intelligent system for healthcare analysis using ridge Adaline stochastic gradient descent classifier, Improve the number of predicted protein functional modules [33], (3) Use samples based deep neural network model for classification of balanced multi-model stroke dataset to predict protein functional modules [34]. (4) Use a novel approach to using big data and IOT for improving the efficiency of *m*-health systems to improve the efficiency of the algorithm [35].

Acknowledgements

2020 Guangxi University Young and middle-aged teachers' basic scientific research ability improvement project No. 2020KY64015; Professor cultivation project of Nanning University in 2020 No. 2020JSGC03, university level scientific research team project of Nanning University No. 2018KYTD01.

References

1 Huo, S.C. and Li, Y. (2019) 'A review of clustering methods for protein functional module detection', *Comput. Eng. Appl.*, Vol. 55, No. 08, pp.17–26.

2 Bano, R. and Rao, K. (2015) 'Graph based gene/protein prediction and clustering over uncertain medical databases', *J. Theor. Appl. Inf. Technol.*, Vol. 82, No. 3, pp.347–352.

3 Ji, J., Zhang, A., Liu, C., Quan, X. and Liu, Z. (2014) 'Survey: functional module detection from protein-protein interaction networks', *IEEE Trans. Knowl. Data Eng.*, Vol. 26, No. 2, pp.261–277.

4 Asur, S., Ucar, D. and Parthasarathy, S. (2007) 'An ensemble framework for clustering protein–protein interaction networks', *Bioinformatics*, Vol. 23, No. 13, pp.i29–40.

5 Javed, A.R., Sarwar, M.U., Beg, M.O., Asim, M., Baker, T. and Tawfik, H. (2020) 'A collaborative healthcare framework for shared healthcare plan with ambient intelligence', *Hum.-Centric Comput. Inf. Sci.*, Vol. 10, pp.1–21.

6 Sarwar, M.U. and Javed, A.R. (2019) 'Collaborative health care plan through crowdsource data using ambient application', *2019 22nd International Multitopic Conference (INMIC), Islamabad, Pakistan*, pp.1–6, doi: 10.1109/INMIC48123.2019.9022684.

7 Javed, A.R., Fahad, L.G., Farhan, A.A., Abbas, S., Srivastava, G., Parizi, R.M. and Khan, M.S. (2021) 'Automated cognitive health assessment in smart homes using machine learning', *Sustain. Cities Soc.*, Vol. 65, p.102572.

8 Aldeco, R. and Marín, I. (2010) 'Jerarca: efficient analysis of complex networks using hierarchical clustering', *Plos One*, Vol. 5, No. 7, pp.11585–11591.

9 Abeysirigunawardena, S.C., Kim, H.J., Lai, J., Ragunathan, K. and Rappé, M.C. (2017) 'Evolution of protein-coupled RNA dynamics during hierarchical assembly of ribosomal complexes', *Nat. Commun.*, Vol. 8, No. 1, pp.492–500.

10 Yao, X.H., Yan, J.W., Liu, K.F., Kim, S.G. and Nho, K.S. (2017) 'Tissue-specific network□base genome wide study of amygdala imaging phenotypes to identif functional interaction modules', *Bioinformatics*, Vol. 33, No. 20, pp.3250–3257.

11 Lin, X.L., Yang, H.Y. and Ye, J. (2015) 'Identification of hot regions in protein-protein interactions based on SVM and DBSCAN', *Intelligent Computing Theories and Methodologies. ICIC 2015. Lecture Notes in Computer Science*, Vol. 9226, pp.390–398.

12 Lei, X., Fang, M., Wu, F.X. and Chen, L.N. (2018) 'Improved flower pollination algorithm for identifying essential proteins', *BMC Syst. Biol.*, Vol. 12, p.46.

13 Ahmad, J., Javed, F. and Hayat, M. (2017) 'Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods', *Artif. Intell. Med.*, Vol. 78, pp.14–22.

14 Lashkov, A.A., Rubinsky, S.V. and Eistrikh-Heller, P.A. (2019) 'Application of the DBSCAN algorithm to detect hydrophobic clusters in protein structures', *Crystallogr. Rep.*, Vol. 64, No. 3, pp.524–532.

15 Melvin, R.L., Godwin, R.C., Xiao, J., Thompson, W.G., Berenhaut, K.S. and Salsbury Jr., F.R. (2016) 'Uncovering large-scale conformational change in molecular dynamics without prior knowledge', *J. Chem. Theory Comput.*, Vol. 12, No. 12, pp.6130–6146.

16 Iwendi, C., Zhang, Z. and Du, X. (2018) 'ACO based key management routing mechanism for WSN security and data collection', *2018 IEEE International Conference on Industrial Technology (ICIT)*, February, IEEE, pp.1935–1939.

17 Mittal, M., Iwendi, C., Khan, S. and Rehman Javed, A. (2021) 'Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using Levenberg–Marquardt neural network and gated recurrent unit for intrusion detection system', *Trans. Emerg. Telecommuni. Technol*., Vol. 32, No. 6, p.e3997.

18 Iwendi, C., Khan, S., Anajemba, J.H., Mittal, M., Alenezi, M. and Alazab, M. (2020) 'The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems', *Sensors*, Vol. 20, No. 9, p.2559.

19 Ji, J.Z. and Gao, G.X. (2017) 'Detecting functional module method based on cultural algorithm in protein-protein interaction networks', *Journal of Beijing University of Technology*, Vol. 43, No. 1, pp.0013–0021.

20 Rodriguez, A. and Laio, A. (2014) 'Clustering by fast search and find of density peaks', *Science*, Vol. 344, No. 6191, pp.1492–1496.

21 Mehmood, R., Bie, R. and Dawood, H. (2015) 'Fuzzy clustering by fast search and find of density peaks', Proceedings *of the 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things*, IEEE, Piscataway, NJ, pp.258–261.

22 Li, M., Wu, X., Wang, J. and Pan, Y. (2012) 'Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data', *BMC Bioinf.*, Vol. 13, p.109.

23 Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisenberg, D. (2000) 'DIP: the data base of interacting proteins', *Nucleic Acids Res.*, Vol. 28, No. 1, pp.289–291.

24 Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T. (2010) 'Cytoscape 2.8: new features for data integration and network visualization', *Bioinformatics*, Vol. 27, No. 3, pp.431–432.

25 Zhang, Y.J., Lin, H.F. and Yang, Z.H. (2017) 'An uncertain model-based approach for identifying protein complexes in uncertain protein-protein interaction networks', *BMC Genomics*, Vol. 18, No. 7, pp.743–752.

26 Halim, Z., Waqas, M. and Hussain, S.F. (2015) 'Clustering large probabilistic graphs using multi-population evolutionary algorithm', *Inf. Sci.*, Vol. 317, No. 1, pp.78–95.

27 Hu, Q.S. and Lei, X.J. (2015) 'Improved Markov clustering algorithm for PPI network', *Comput. Sci.*, Vol. 42, No. 7, pp.108–113.

28 Lei, X., Li, H. and Zhang, A. (2017) 'iOPTICS-GSO for identifying protein coMLPexes from dynamic PPI networks', *BMC Medical Genomics*, Vol. 10, Suppl. 5, p.80.

29 Chi, X. and Hou, J. (2011) 'An iterative approach of protein function prediction', *BMC Bioinf.*, Vol. 12, No. 1, pp.437–445.

30 Teng, Z., Guo, M. and Liu, X. (2013) 'Revealing protein functions based on relationships of interacting proteins and GO terms', *J. Comput. Biol.*, Vol. 20, No. 4, pp.322–343.

31 Krogan, N., Caney, G. and Yu, H. (2006) 'Global landscape of protein complexes in the yeast Saccharomyces Cerevisiae', *Nature*, Vol. 440, pp.637–643.

32 Chakraborty, C. (2021) 'Performance analysis of compression techniques for chronic wound image transmission under smartphone-enabled tele-wound network', *Research Anthology on Telemedicine Efficacy, Adoption, and Impact on Healthcare Delivery*, IGI Global, pp.345–364.

33 Deepa, N., Prabadevi, B., Maddikunta, P.K., Gadekallu, T.R., Baker, T., Khan, M.A. and Tariq, U. (2021) 'An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier', *J. Supercomput.*, Vol. 77, pp.1998–2017.

34 Bhattacharya, S., Maddikunta, P.K.R., Hakak, S., Khan, W.Z., Bashir, A.K., Jolfaei, A. and Tariq, U. (2020) 'Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset', *Multimed. Tools Appl.*, pp.1–25.

35 Mishra, K.N. and Chakraborty, C. (2020) 'A novel approach towards using big data and IoT for improving the efficiency of m-health systems', *Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare*, pp.123–139.