Based on PPI Network and Deep Learning Predict Protein Function Module Algorithm

Sicong Huo¹, Tao Lu²

¹ Nanning University, Nanning, Guangxi, 530200, China
² Nanning University, Nanning, Guangxi, 530200, China

* Sicong Huo e-mail: 237290097@qq.com

Abstract. The existing Protein-protein interaction (PPI) based prediction algorithm for protein functional modules has the disadvantages of low accuracy and less number of functional modules. This paper presents an algorithm based on PPI network and in-depth learning and prediction of protein functional module. In this method, three important attributes (node location, PPI network structure and core nodes) are integrated to improve the prediction accuracy and increase the number of protein functional modules. In this paper, First use the Local Density-Based Methods (LDBM) improved density clustering algorithm for clustering analysis. After clustering, use the improved to isometric mapping dimension reduction algorithm for principal component analysis (PPIPCA); then use Multi Layer Perceptron (MLP) for training and function module selection. Last Based on PPI network and deep learning based prediction protein function module algorithm (LPMM) compared with other algorithms. On the basis of yeast dip data set, the accuracy, F value, dimensionality reduction rate and functional module of recognition protein of LPMM algorithm and acc-fmd, MCL and mcode algorithm based on PPI network were compared. Experimental results show that LPMM algorithm is superior to other algorithms in accuracy, F value, dimensionality reduction rate, number of recognition modules and so on.

1. Introduction

In recent years, as the bioinformation technology develops, prediction of protein function module has become a research hotspot in the field of bioinformation. At present, many researchers predict protein function modules according to the characteristics of Protein-protein interaction (PPI), and a large number of protein data are mined out. Prediction of protein function module involves many technologies. More and more researchers use the three-dimensional data structure of protein to predict protein function module, and successfully predict more and more new protein function modules. This kind of research direction has become a new hotspot for researchers to predict protein function module. At present, the intelligent algorithm plays an important role in the prediction of protein function module. This kind of algorithm has good nonlinear fitting ability. It can map the linear nodes of PPI network to any complex nonlinear relationship. The learning rules are simple, the programming implementation is easy to achieve, and it has strong robustness, memory ability and global approximation ability, which solves the problem The basic problems of protein function module prediction. However, the new algorithm still has some shortcomings, such as accuracy, low F value, and the number of protein functional modules. In view of these shortcomings, researchers use the characteristics of PPI network three-dimensional data structure to predict protein function module [1].
The reliability of protein function module based on PPI network is high. This paper proposes an algorithm based on PPI network and deep learning to predict protein function module, which integrates three important information of PPI network (protein core node, protein each node location, PPI network structure) as the important attribute of protein function module, proposes an Local Density-Based Methods (LDBM) for clustering analysis. After clustering, use the improved to isometric mapping dimension reduction algorithm for principal component analysis (PPIPCA), and then the Multi Layer Perceptron (MLP) is used to train the protein function module prediction method. Based on PPI network and deep learning based prediction protein function module algorithm (LPMM) has higher accuracy and F value than other algorithms.

2. Algorithm of protein function module based on PPI network and deep learning

PPI network is usually represented as an undirected graph G (V, E) \(^1\), where V = \{V_1, V_2, V_3, \ldots V_N\}, expressed as node set, E = \{E_1, E_2, E_3, \ldots E_M\} is expressed as a set of connecting edges of each node. In this paper, a protein function prediction algorithm is based on LPMM and PPI network proposed. As shown in Figure 1, LPMM algorithm is divided into four parts: 1. Data input; 2. Training model; 3. Comparison with known protein function module; 4. Data output. First of all, the input data is the known protein function module interaction annotation term data and the main characteristics of PPI network (protein core node, protein each node location, PPI network structure); the input data is normalized; secondly, the multi-layer sensor (MLP) is used for training; then two classifiers are used to classify the predicted protein function module; finally, the function module is selected for output.

![Fig 1 LPMM algorithm flowchar](image)

2.1. Extract the main attributes of PPI network

The research shows that the closer the two core nodes in PPI network are, the greater the probability of predicting the protein function modules \(^2\). Using the characteristics of PPI network, we can first mine some known or unknown protein function modules by clustering, and then use the main characteristics of clustering to predict the protein function modules. The noise of this method is small and robust.

Density clustering algorithm is often used to predict the function modules of PPI network. The traditional density algorithm needs human intervention to determine the protein core nodes. In this
paper, an improved density clustering algorithm is proposed to predict protein function modules. According to formula (1) and formula (2), the local density is calculated and the central node is determined. This is the result of the first clustering:

\[ \rho_i = \sum_j x(d_{ij} - d_c) \]  
\[ \delta_i = \min_{j \in P_i} (d_{ij} - d_c) \]

The remaining undivided node sets were calculated again to obtain the maximum local density and determine the next core node of PPI network, and then clustering analysis was conducted. Repeat the above process. Until all node data are clustered or sparse PPI network nodes are treated as noise, a threshold condition is finally set to determine whether the remaining data nodes are regarded as noise. The specific steps of the algorithm are shown in the figure below.

2.2. Principal component analysis
As mentioned above, three main attributes (core node, core node location, PPI network structure) in PPI network play a key role in the prediction of protein function modules. These three main attributes are used to predict protein function modules to improve accuracy and reduce data noise. In this paper, MLP is used in training. If the clusters are trained, the high dimension of clusters will lead to high noise and affect the prediction results. Considering the correlation of protein functional attributes, it is necessary to analyze the clusters after clustering. In order to maintain the consistency of data structure before and after dimensionality reduction, an improved isometric mapping dimensionality reduction algorithm is proposed. The basic idea is as follows: according to the principle of local linearity, the matrix is established by clustering clusters; then the minimum distance between any cluster nodes is calculated, which is expressed as the geodesic distance; finally, the low-dimensional representation of high-dimensional data is obtained by using multiple dimensional scaling (MDS) algorithm.

The specific process of the improved isometric mapping dimensionality reduction algorithm is as follows: (1) construct the adjacency matrix. Firstly, the clustering cluster is regarded as the neighborhood point. The method of selecting neighborhood points: \( \varepsilon \) is the neighborhood radius, and when \( \|x_i - x_j\| < \varepsilon \), \( X_j \) is the neighborhood point of \( X_i \). Establish weighted adjacency matrix:
Calculate the shortest path. If Xi and XJ are domain points, the geodesic distance can be replaced by the Euclidean distance s(i, j) between them. Otherwise, the algorithm sets the starting geodesic distance $S_{C}(i, j) = \infty$, and constructs the shortest path as follows:

$$s_{i}(x_{i}, x_{j}) = \min_{k=1-n} \{s_{i}(x_{i}, x_{k}), s_{j}(x_{k}, x_{j}) + s_{k}(x_{k}, x_{j})\}$$ (4)

The above methods, the geodesic distance in the neighborhood and the geodesic distance outside the neighborhood are obtained respectively, and the shortest path moment representing the geodesic distance between sample points is obtained. $S_{C} = [S_{C}(i, j)]_{n \times m}$ shortest path distance converges to geodesic distance $S_{C}$, and then w matrix is applied to MDS algorithm for dimension reduction.

In order to improve the quality of dimensionality reduction, the geodesic is modified as follows and calculated by formula (5):

$$S_{C} = \left[ s_{C}(x_{i}, x_{j}) \right]_{N \times N}, i, j = 1, 2, \cdots, N$$ (5)

Where $S_{C}(X_{i}, X_{j})$ is determined by the following formula (6):

$$s_{i}(x_{i}, x_{j}) = \min \left\{ \sum_{k=1}^{N} n(x_{i}, x_{k}) \sum_{l=1}^{N} s(x_{k}, x_{j}) + s(x_{k}, x_{j}) + \sum_{m=1}^{N} d(x_{k}, x_{j}) \right\}$$ (6)

The distance s shown in formula (7) is Euclidean distance, and the core matrix (kernel matrix) $k$ is calculated by geodesic distance:

$$K = HS_{C}^{2}H$$ (7)

Formula (7) matrix $H$ is calculated by the following formula:

$$H = I - \frac{1}{N} e_{N} e_{N}^{T}$$ (8)

Where $e_{n} = [0 \ 1]^{T} \in w_{n}$. The matrix $K$ is eigendecomposed to get the maximal n values $\lambda_{1} \ \lambda_{2} \ \lambda_{3} \ \cdots \ \lambda_{n}$, and the corresponding eigenvector $\omega_{1} \ \omega_{2} \ \omega_{3} \ \cdots \ \omega_{n}$. Matrix composed of $\omega_{n}$, using matrix $\omega = \{\omega_{1}, \ \omega_{2}, \ \omega_{3} \ \omega_{n}\}$ $\omega \rightarrow x_{ij}$ gives the corresponding low dimensional data $X_{ij}$:

$$X'_{ij} = \left\{ x_{ij} \mid x_{ij} \in R^{k}, j = 1, 2, \cdots, N \right\}$$ (9)

2.3. Multi sensor training and functional protein selection

The core cluster of PPI network after clustering and dimensionality reduction reflects the importance of protein functional modules in PPI network. The higher the importance of core cluster nodes, the more functional properties of protein.\(^{[4]}\) In this paper, the core cluster after dimensionality reduction is regarded as a feature matrix, which is combined with the matrix composed of the known main components of protein function module attributes, and a new feature matrix is obtained. In order to improve the quality of training, the characteristic matrix is standardized by min max to make each element data linear transformation.

In this paper, MLP with single hidden layer is used as classifier. Compared with the traditional classifier, MLP classifier is more flexible, which overcomes the weakness that perceptron can't recognize the linear non separable data. It uses PPI network to form the model of node number, activation function and loss function of cluster, and uses hidden layer to mine the complex relationship between the predicted protein function module and the known protein function module. In PPI network, the characteristic vectors of the data terms of the known protein function modules are used as the input, and the annotation of the known protein function modules is used as the output to establish the mapping relationship between the multi features and multi functions in the protein function modules.

3. Experimental data set and evaluation index

In order to verify the effectiveness of the LPMM algorithm proposed in this paper, PPI network
interaction data is relatively complete and reliable as the experimental data. In the experiment, firstly, we download PPI network data from dip database, use uniprotkb / Swiss prot to perform ID conversion on proteins in PPI network, process and filter repeated interaction and untranslatable protein terms through serial affinity purification, and then obtain dip term number interpro term number according to uniprotkb / Swiss prot number of each protein this paper, precision, recall and F-measure and OS (a, b) are the indexes to measure the prediction effect of the algorithm. The formula is as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{10}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{11}
\]

\[
\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}
\]

\[
\text{OS}(A, B) = \frac{|A \cap B|}{|A| \times |B|} \tag{13}
\]

3.1. Experimental data analysis

In order to analyze the protein function modules detected by LPMM algorithm on PPI network in detail, based on the PPI network data provided by dip database, LPMM algorithm is compared with ACC-FMD\(^9\), MCL\(^6\), MCODE\(^7\) algorithm.

The number of predicted protein functional modules (PM) of LPMM algorithm compared with other three algorithms is shown in Table 1.

| Approaches | PM | TP | FULL |
|------------|----|----|------|
| LPMM       | 392| 253| 19   |
| ACC-FMD    | 374| 101| 9    |
| MCL        | 126| 52 | 5    |
| MCODE      | 254| 129| 8    |

The number of significant protein function modules (SC) mined by LPMM algorithm and other three algorithms is shown in Table 2.

| Approaches | PM | SC  | TP/PM      |
|------------|----|-----|------------|
| LPMM       | 392| 153 | 39.03%     |
| ACC-FMD    | 374| 85  | 22.73%     |
| MCL        | 126| 36  | 28.57%     |
| MCODE      | 254| 98  | 38.58%     |

Table 3 shows the comparison of dimension reduction rate \(1-N / M\) between lpmm algorithm and other three algorithms.

| Approaches | M   | N   | 1-N/M    |
|------------|-----|-----|----------|
| LPMM       | 905 | 223 | 75.36%   |
| ACC-FMD    | 823 | 488 | 40.70%   |
| MCL        | 764 | 346 | 54.71%   |
| MCODE      | 819 | 459 | 43.96%   |

The comparison between lpmm algorithm and other three algorithms, precision, recall, F-measure, shown in Table 4.
Table 4 Efficiency analysis of LPMM algorithm and three other algorithms

| Approaches | Precision | Recall | F-Measure |
|------------|-----------|--------|-----------|
| LPMM       | 0.381     | 0.352  | 0.366     |
| ACC-FMD    | 0.231     | 0.187  | 0.207     |
| MCL        | 0.129     | 0.404  | 0.196     |
| MCODE      | 0.328     | 0.268  | 0.295     |

The experimental results show that 253 of the protein function modules (PM) mined by LPMM algorithm are successfully matched by known protein function modules, which indicates that LPMM algorithm can recognize more protein function modules. The dimensionality reduction rate (1 - N / M) is 75.36%, which is higher than other three algorithms. As shown in Table 4, the LPMM algorithm reflects the efficiency of the LPMM algorithm in the evaluation and prediction of the index accuracy of the protein function module (precision) 0.381, recall (recall) 0.352 and F-measure (F-measure) 0.366.

4. Conclusion
In the face of the continuous growth of PPI network data, how to efficiently predict protein functional modules is an important research issue in modern bioinformatics. It is a popular method to predict protein function module based on PPI network. This kind of method has low cost and high efficiency, but the data obtained is noisy. This paper proposes an LPMM algorithm for this problem. Experimental results show that LPMM algorithm can effectively predict protein function modules. In the next step, Researcher can study predict protein function modules from the following directions. How to predict the protein function module based on PPI network and swarm intelligence optimization algorithm, such as Bayesian intelligence algorithm, and neural network intelligence algorithm Prediction of protein functional modules.

Appendices
This study is supported by the National Natural Science Foundation of China under Grant Nos. 61741203,61866006; Guangxi Natural Science Foundation (2016GXNSFAA380243,); Guangxi Nanning Science and Technology Development Planning Project(20181015-5)

References
[1] Huo Sicong, Li Ying. A review of clustering methods for protein functional module detection [J]. Computer engineering and application, 2019,55 (08): 17-26.
[2] Dynamic weighted PPI network complex mining based on ant colony clustering [J / OL]. Computer application research: 1-10 [2019-06-13]. https://doi.org/10.19734/j.issn.1001-3695.2018.07.0518.
[3] Chen B, Fan W, Liu J, et al. Identifying protein coMLPexes and functional modules--from static PPI networks to dynamic PPI networks.[J]. Briefings in Bioinformatics, 2014, 15(2):177.
[4] Liu Wei, Liu Shang, Zhou Xuan. Study on BP neural network sub batch learning method[J]. Journal of intelligent systems, 2016, 11 (2): 226-232.
[5] Liu Hongxin. Research on PPI network functional module detection algorithm based on ant colony clustering mechanism [D]. Beijing University of technology, 2013
[6] Hu Qingsheng, Lei Xiujuan. Improved Markov clustering algorithm for PPI network [J]. Computer science, 2015, 42 (7): 108-113
[7] Lei X, Li H, Zhang A, et al. iOPTICS-GSO for identifying protein coMLPexes from dynamic PPI networks[J]. Bmc Medical Genomics, 2017, 10(Suppl 5):80
[8] Zhang X F, Ouyang L, Zhu Y, et al. Determining minimum set of driver nodes in protein-protein interaction networks[J]. Bmc Bioinformatics, 2015, 16(1):146.
[9] Song J,Singh M. How and When should Interactome-derived Clusters be Used to PredictFunctional Modules and Protein Function[J].Bioinformatics,2009,25(29):3143-3150.