A Key Protein Recognition Algorithm Based on High-order Complex Protein Network

Xiaojing Wang

Information Management Center
Hohhot Vocational College
Hohhot 010010, China
E-mail: 200800553@hhvc.edu.cn

Received: October 01, 2021 **Accepted: March 17, 2022**

Published: June 30, 2022

Abstract: *Although the existing protein recognition methods have improved the recognition accuracy of key proteins to a certain extent, they have ignored the biological features of the proteins. In view of this shortcoming, this paper constructed a high-order dynamic complex protein network for key protein recognition. At first, this paper presented a method for feature selection and candidate set evaluation of complex protein network; a weighted network was constructed based on the obtained topological features of the complex protein network and the semantic similarity of protein gene ontology annotations. Then, this paper proposed an algorithm for recognizing key proteins in high-order dynamic protein network based on a Fruit fly optimization algorithm. At last, the effectiveness of the proposed model was verified by experimental results.*

Keywords: *Complex protein network, Dynamic network, Key protein recognition, Fruit fly optimization algorithm (FOA).*

Introduction

With the implementation and promotion of the human genome project, the sequencing of human genome has been completed, which is taken as a symbol of the arrival of the post genome era [1, 5, 6, 9, 10, 15, 23]. Although the genome sequencing project has finished already, still, we haven't fully figured out the internal mechanism of genome sequence [3, 4, 11, 13, 17]. Proteins are the products of gene expression; the study of proteomics enables us to probe deep into the internal relations of functional genomes [2, 22]. Every life activity needs to be completed by the cooperation of multiple kinds of proteins. According to the function of proteins in life activities, they can be divided into key proteins and non-key proteins [14, 16, 19]. The recognition of key proteins not only helps us understand the law of cell activities, but also provides a solid theoretical basis for the research on disease pathogenesis and the development of corresponding medicines.

Aiming at the problem of low recognition degree of essential proteins based on topological parameters, Huang et al. [8] analyzed the correlations between the essentiality of proteins and their main topological parameters, and discussed the nature of the essentiality-judgment ability of parameters; then they made use of such correlations among parameters to obtain the mutual information of essential nodes contained in these parameters and put forward the construction method of parameter combinations. Protein fold recognition is one of the important steps in protein structure prediction. Yan et al. [20] combined two main computational approaches: the template-based method based on the alignment scores between query-template protein pairs, and the machine learning method based on feature representation and machine learning classifier, they integrated the advantages and disadvantages of the two methods, and proposed more accurate predictors for protein fold recognition. The tertiary structure of proteins is

determined by the amino acid sequence in the process of protein folding, and it plays an important role in protein functions. Protein fold recognition is one of the hotspots in the study of bioinformatics. Hekmatnia et al. [7] proposed a feature selection method based on Map Reduce framework and Vortex Search Algorithm, and experimental results proved that their method had greatly improved the prediction accuracy. Lei and Zhang [12] proposed a logistic regression algorithm for identifying candidate disease genes based on reliable protein-protein interaction network, and achieved good recognition effect. Existing methods mostly recognize protein functional complexes from the protein-protein interaction networks at a good level, the applicability of advanced graph network methods has not yet been fully studied. Zaki et al. [21] proposed various graph convolutional network methods to improve the detection of protein complexes, they developed a neural overlapping community detection model to cluster the nodes using a complex affiliation matrix, and they found that the performance of the algorithm was significantly better than the previous advanced methods. Khattak et al. [18] applied the protein-protein interaction network to the analysis and exploration of genes related to oral cancer diseases; the proposed technology is fully interactive and can more accurately and effectively analyze the data of oral cancer diseases.

Although existing protein recognition methods have improved the recognition accuracy of key proteins to a certain extent, they generally have ignored the biological features of proteins, and the real proteins are constantly changing in the cell cycle. In order to make up for the shortcomings of existing methods in key protein recognition, this paper constructed a high-order dynamic complex protein network for key protein recognition. The content of this paper mainly contains the following several aspects: 1) the proposal of a method for feature selection and candidate set evaluation of the complex protein network; 2) the construction of a weighted network based on the obtained topological features of the complex protein network and the semantic similarity of protein gene ontology annotations; 3) effectively recognizing key proteins in the high-order dynamic protein network based on Fruit fly optimization algorithm (FOA); 4) verifying the effectiveness of the constructed model using experimental results.

Network feature selection and candidate set evaluation

A gene ontology annotation is a description of the function of a specific gene. Each annotation is composed of a gene and the descriptive vocabularies of the corresponding function, which can facilitate the various studies of researchers. Fig. 1 shows the structure of a gene ontology annotation. As can be seen from the figure, the semantic concept of function description that is closer to the root node contains less information, while the semantic concept of function description that is farther away from the root node contains more information.

Cosine similarity was adopted to calculate the similarity of two gene ontology annotation vectors in the target protein annotation text. Fig. 2 gives a diagram of cosine similarity.

Suppose: the gene ontology annotation vector n is represented by $[u_1, v_1]$; vector m is represented by $[u_2, v_2]$, then, based on the cosine value of angle ω between n and m, the cosine similarity could be calculated:

$$cos\omega = \frac{u_1u_2 + v_1v_2}{\sqrt{u_1^2 + v_1^2} \times \sqrt{u_2^2 + v_2^2}} \quad (1)$$

Suppose, in a m-dimensional space, X is represented by $[X_1, X_2, \ldots, X_m]$, Y is represented by $[Y_1, Y_2, \ldots, Y_m]$, then the cosine similarity can be calculated by Eq. (2):

$$cos\omega = \frac{\sum_{i=1}^{m}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{m}(X_i)^2} \times \sum_{i=1}^{m}(Y_i)^2} = \frac{X \cdot Y}{|X| \times |Y|} \quad (2)$$

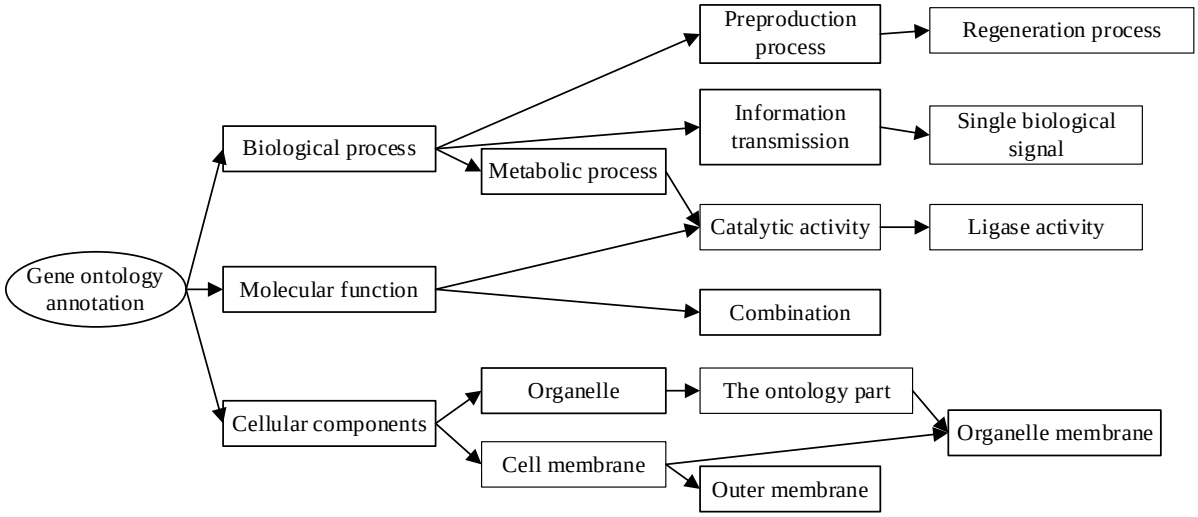


Fig. 1 Structure of a gene ontology annotation

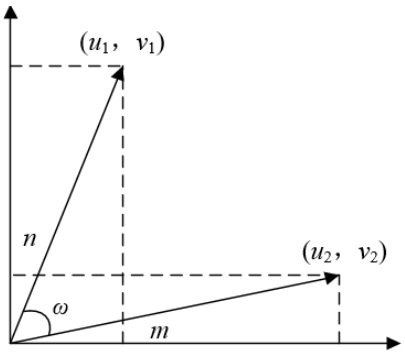


Fig. 2 A diagram of cosine similarity

Based on cosine similarity, the complex protein network model could be constructed. For example, count the word frequency of annotation text of the protein pairs to obtain corresponding high-frequency word sets, that is, to construct the complex protein network model based on the similarity of high-frequency words.

Suppose: B represents the set of protein nodes, R represents the set of node connection edges, then the constructed complex protein network can be defined as $H = \{B, R\}$. The unweighted complex protein network can be described by the adjacent matrix $X = \{x_{ij}\}$, and the information about the connectivity between key protein nodes i and j is stored by the matrix element x_{ij}. If nodes i and j are connected, x_{ij} is equal to 1; otherwise, if nodes i and j are not connected, x_{ij} is equal to 0. For the constructed complex protein network based on annotation text, this paper listed a few parameters to characterize its topological structure:

If the access between two protein nodes can be completed through the shortest path, it can be judged that the two nodes are related, and the relevance between the two nodes can be described

by an intermediary value. Suppose: among the shortest paths between protein nodes *a* and *b*, the number of paths passing through node *i* is represented by m_{abi}, and the total number of shortest paths connecting *a* and *b* is represented by m_{ab}, then Eq. (3) gives the calculation equation of the intermediate value:

$$Y_i = \frac{1}{M^2} \sum_a \sum_b \frac{m_{abi}}{m_{ab}} \tag{3}$$

Suppose: degree *l* represents the number of edges connected to a protein node, it satisfies $l_i = \sum_j x_{ij}$. The analysis of protein annotation text requires degree-related features, such as the average node degree defined as Eq. (4) and the standard deviation of neighbor node defined as Eq. (5):

$$l_i^{(m)} = \frac{1}{l_i} \sum_j x_{ij} l_j \tag{4}$$

$$\Delta l_i^{(m)} = \left[\frac{1}{l_i} \sum_j x_{ij} \left(l_j - \frac{1}{l_i} \sum_n x_{in} l_n \right)^2 \right]^{1/2} \tag{5}$$

Suppose: *r* represents the total number of protein node edges, then, the symbiosis between protein nodes, namely the analysis of protein node types can be measured by Pearson correlation coefficient shown in Eq. (6):

$$s = \frac{r^{-1} \sum_{j>i} l_i l_j x_{ij} - \left[r^{-1} \sum_{j>i} \frac{1}{2} \left(l_i + l_j \right) x_{ij} \right]^2}{r^{-1} \sum_{j>i} \frac{1}{2} \left(l_i^2 + l_j^2 \right) x_{ij} - \left[r^{-1} \sum_{j>i} \frac{1}{2} \left(l_i + l_j \right) x_{ij} \right]^2} \tag{6}$$

In the gene ontology annotation neighborhood network of annotation text of the target protein, how gene ontology annotations with different frequencies appearing as neighbor nodes needs to be quantified using the Pearson correlation coefficient.

The local node density of protein node neighborhood can be described by the clustering coefficient, and the number of triangles between three surrounding nodes close to the target protein node can be used for equivalent calculation:

$$D = 3 \sum_{l>j>i} x_{ij} x_{il} x_{jl} \left[\sum_{l>j>i} x_{ij} x_{il} + x_{ji} x_{jl} + x_{li} x_{lj} \right] \tag{7}$$

After obtaining the eigenvalues of topological structure of the complex protein network, the evaluation of vocabulary patterns in the candidate set can be completed based on the eigenvalues of gene ontology annotations of the corresponding protein annotation text. Based on the judgement principle of whether the evaluation score is the highest or not, this paper performed recognition on the most important gene ontology annotation in a certain vocabulary pattern, that is, most important gene ontology annotation has the strongest ability to express this vocabulary pattern. Suppose: LM (t_i) represents the evaluation score of vocabulary pattern

before introducing the gene ontology annotation importance judgment, $MA(Q_i)$ represents the largest evaluation score of gene ontology annotations. By introducing the evaluation score of gene ontology annotation of target protein annotation text into the evaluation of vocabulary pattern, we can get a new evaluation equation as follows:

$$LM'(t_i) = LM(t_i) * MA(Q) \tag{8}$$

The new scores of protein pairs can be obtained by combining the original scores of each protein pair with the eigenvalues of the complex network. Suppose: $PJ(e)$ represents the average topological attribute value of the complex protein network, $LM(e)$ represents the original score of a target protein pair, then the calculation equation of the new score $LM^*(e)$ is:

$$LM^*(e) = LM(e) * PJ(e) \tag{9}$$

The above evaluation method can realize the protein interaction recognition algorithm, thereby better evaluating the proteins in the candidate set and making the proteins added to the candidate set more reliable.

Construction of the Complex Protein Network

Topological and biological feature extraction

This section elaborates on the construction process of the complex protein network. First, the topological features of the network used for evaluating the tightness of protein-protein interactions were calculated, including the clustering coefficients of protein nodes and edges.

The clustering coefficient of protein nodes can be defined as the degree of nodes and the number of edges connecting other nodes, to a certain extent, it can represent the tightness of connection between a certain protein node and other adjacent protein nodes. Here, it is defined that the degree of protein a is represented by l_a, which means that there are l_a edges connecting with other protein nodes through a, the set of its neighborhood protein nodes is represented by $M(a)$. At this time, the number of effective interaction edges of the sub-network constituted of a and $M(a)$ can be represented by $R_{M(a)}$, then Eq. (10) gives the calculation equation of the clustering coefficient GE_a of a:

$$GE_a = \frac{2 \times R_{M(a)}}{l_a \times (l_a - 1)} \tag{10}$$

Here, it is defined that the edge between proteins a and b is represented by (a, b), the number of protein nodes jointly connected by a and b is represented by C_{ab}, and the degrees of a and b are respectively represented by l_a and l_b, then Eq. (11) gives the calculation equation of the clustering coefficient $W(a, b)$ of edges:

$$GW(a,b) = \frac{C_{ab}}{min(l_a - 1, l_b - 1)} \tag{11}$$

If a and b are not connected, namely there is no interaction between a and b, then $GW(a, b)$ is equal to 0.

The possibility that a protein belongs to a similar protein complex can be evaluated based on the similarity of protein gene ontology annotation information. The similarity between gene products based on gene ontology annotation is regarded as the semantic similarity between gene ontology annotations. The maximum information volume of the most informative common ancestor node of the protein nodes, namely the lowest common ancestor, was equivalent to the semantic similarity between gene ontology annotations. Suppose: $t_{ne}(a, b)$ represents the number of occurrences of the lowest common ancestor of nodes a and b in the annotation text database, then Eq. (12) gives the calculation equation of information volume AX:

$$AX(e, p) = -log[t_{ne}(a, b)] \tag{12}$$

Suppose: δ_{LC} represents the gene ontology annotation of the lowest common ancestor of gene ontology annotation information δ_i and δ_j, then the semantic similarity between gene ontology annotations can be described by Eq. (13):

$$sim(\delta_i, \delta_j) = max[AX(\delta_{LC})] \tag{13}$$

Construction of the weighted network

When only the topological features are used to weight the complex protein network, the low quality of network data and the lack of protein complexes will affect the reliability of the evaluation of protein interactions. This paper constructed the weighted network based on the obtained topological features of the complex protein network and the semantic similarity of protein gene ontology annotations.

$W(a, b)$ was re-defined based on the Jaccard similarity coefficient:

$$JCD(a, b) = \left| \frac{a \cap b}{a \cup b} \right| \tag{14}$$

The value range of $JCD(a, b)$ is [0, 1]. Combining with the Jaccard similarity coefficient, the improved edge clustering coefficient in $H = \{B, R\}$ can be defined as GW, thereby obtaining the calculation result $GW(a, b)$ of the similarity of topological structure of the complex protein network. Suppose: $M(a)$ and $M(b)$ represent the sets of other surrounding proteins that connect to the edges of protein node a and b, then there is:

$$GW(a, b) = \left| \frac{M(a) \cap M(b)}{M(a) \cup M(b)} \right| \tag{15}$$

Where, the numerator is the number of neighbor protein nodes commonly owned by a and b, and the denominator is the total number of neighbor protein nodes connected with a and b. After the similarity of topological structure of the complex network has been defined as GW, the relationships between a protein node and other nodes in its neighborhood are no longer equal, and this node will be more inclined to neighbor nodes that are closely connected with it. In order to effectively reduce the impact of the same protein annotation text on the calculation results of the similarity of gene ontology annotations and improve calculation accuracy, this paper defines semantic similarity as the proportion of semantic information volume of gene ontology annotations in the annotation text set with the largest volume of corresponding semantic information in the complex network. Suppose: proteins a and b are respectively

annotated by gene ontology annotation sets X_1 and Y_1; E_1 and E_2 respectively represent the annotation text sets of the semantic information of a and b; E_1 and E_2 contain the annotation text set $AX_{max}(E)$ with the largest semantic information volume; $t(\delta_i)$ represents the number of occurrences of a gene ontology annotation δ_i in the specified annotation database; $AX(\delta)$ represents the semantic information volume of δ_i, then, the semantic similarity $SS(a, b)$ between a and b is defined by Eq. (16):

$$SS(a,b)=\frac{\sum_{\delta_i\in X_1\cap Y_i} AX(\delta_i)}{AX_{max}(E)}=\frac{\sum_{\delta_i\in X_1\cap Y_i} -logt(\delta_i)}{max\{AX(E_1),AX(E_2)\}} \tag{16}$$

The specific construction process of the weighted network based on obtained topological structure features of the complex protein network and the semantic similarity of protein gene ontology annotations is as follows: at first, the topological structure similarity of the network and the semantic similarity of gene ontology annotations were summed and bisected, and the calculation equation of the similarity $S(a, b)$ of protein nodes a and b corresponding to any connecting edge in the complex network is given by Eq. (17):

$$S(a,b)=\frac{\sum GW(a,b)+\sum SS(a,b)}{2} \tag{17}$$

The value range of $S(a, b)$ in above equation is [0, 1]. In the obtained weighted protein network, the weight value of the interaction edge of proteins can represent the similarity between the two, namely the tightness between them.

Recognition of key proteins in high-order dynamic protein network

Model construction

In order to improve the accuracy and efficiency of key protein recognition in the complex protein network, this paper fully considered the time series of the constructed protein network, and used FOA to conduct effective recognition of key proteins in the high-order dynamic protein network.

Firstly, the protein's active period was introduced into the constructed complex protein network to generate a high-order dynamic protein network model. Fig. 3 gives a diagram of the constructed high-order dynamic protein network.

This paper characterized the threshold of gene information volume according to the 3-σ principle. If a protein in the network is active, its corresponding gene information volume value is greater than the preset standard threshold of the 3-σ principle. On the contrary, it is less than that.

Suppose: $\lambda(a)$ and $\varepsilon(a)$ represent the mean and variance of gene information volume of protein a within time period $[1, \psi]$, then Eq. (18) gives the calculation equation of threshold $FZ(a)$:

$$FZ(a)=\lambda(a)+3\varepsilon(a)\left(1-\frac{1}{1+\varepsilon^2(a)}\right) \tag{18}$$

Due to the periodicity of gene information data generation, the gene information volume value at a certain time is defined as the mean of gene information volume in three cycles. Suppose: $\psi(i)$ represents the gene information volume at time moment t, then there is:

$$QD(i)=\frac{\psi(i)+\psi(i+10)+\psi(i+20)}{3},\ (i \in [1, 10]) \tag{19}$$

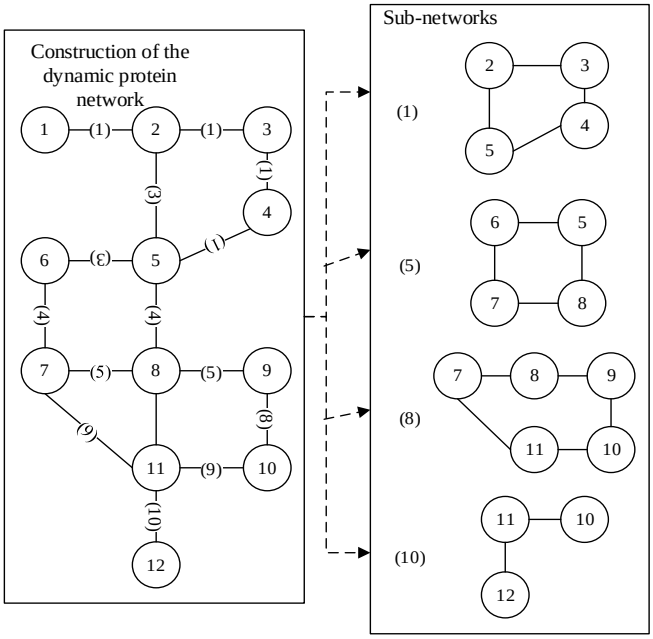


Fig. 3 Diagram of the constructed high-order dynamic protein network

At a fixed time, if an interacting protein pair is active, then the connecting edge of the proteins is also active. The entire protein network can be divided into several subnetworks based on the active state of the proteins, here the number of subnetworks takes 10.

For the constructed weighted high-order dynamic protein network, in order to obtain effective topological and biological information of the network and get a new centrality strategy, the centrality score $CH(a)$ of protein a can be calculated using the equation below:

$$CH(a)=\sum_{a\in M(a)} GW(a,b)\times S(a,b) \tag{20}$$

Since there is a certain probability for each protein to generate 10 dynamic protein subnetworks, when calculating $CH(a)$, the frequency of proteins appearing in the subnetworks should be fully considered. Suppose: $CH^i(b)$ represents the centrality score of protein a at time moment i, $CS(b)$ represents the number of occurrences of protein a in the dynamic protein subnetworks, then, there is:

$$W_{CH}(a)=\frac{\sum_{i=1}^{10} CH^i(a)}{CS(a)} \tag{21}$$

According to above equation, if protein a does not appear in subnetworks at time moment i, then $CH^i(b)$ is equal to 0.

Key protein recognition

For different swarm intelligence algorithms, the performance varies greatly. FOA has the merits of powerful global searching ability, small computation load, and low complexity. The fruit fly swarm has the swarm intelligence of fast searching, that is, visual search enables the fruit flies to quickly locate the optimal location, and then the location information is spread to the entire swarm, therefore, FOA is very good at searching for global optimal. The olfactory searching mechanism shows that fruit fly individuals have a certain ability to jump out of local optimal locations, then combining with visual searching, the fruit fly swarm can migrate locations gradually and update the information of current optimal location.

This paper effectively recognized key proteins in the high-order dynamic protein network based on FOA and associated the process of fruit fly searching for the optimum with the process of key protein recognition. Fig. 4 shows the flow of the key protein recognition algorithm. The key protein candidate set and the sequence number of candidate proteins were associated with each fruit fly individual and its corresponding location.

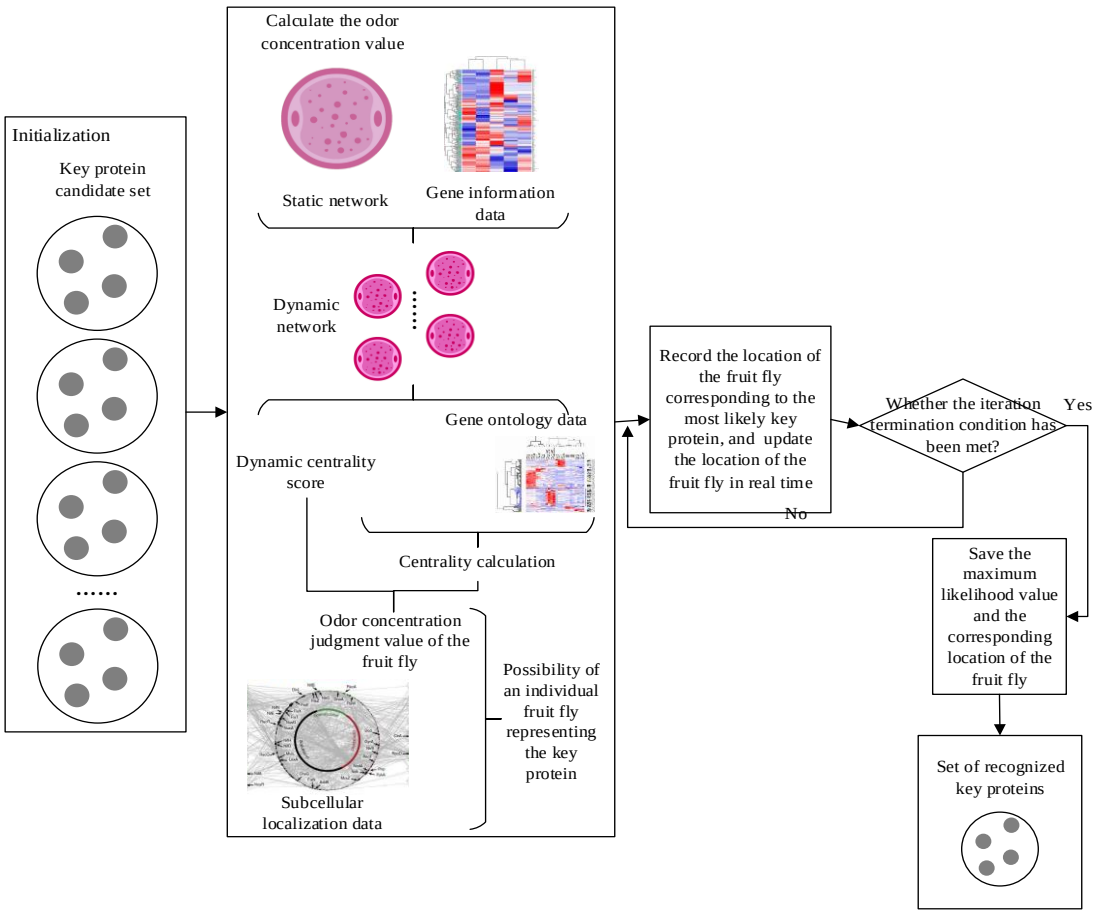


Fig. 4 Flow of the key protein recognition algorithm

Suppose: $W_{ZXD}(a_j)$ represents the average score of the dynamic local connectivity of the j-th protein in q key protein candidate sets, $W_{CH}(a_j)$ represents the corresponding dynamic network topology centrality score, in order to comprehensively evaluate the topological features of the constructed weighted high-order dynamic protein network, this paper combined W_{ZXD} (the parameter describing the modularity degree of the network) with W_{CH} (the parameter describing the centrality of the new high-order dynamic protein network), and Eq. (22) gives

the equation for calculating the odor concentration judgment value $OC(i)$ of fruit fly at a certain location during the execution of the algorithm:

$$OC(i) = \sum_{j=1}^{q} \left(W_{ZXD}(\lambda_j) + W_{CH}(\lambda_j) \right) \tag{22}$$

The accurate calculation of the topological and biological features of the network is the basis of key protein recognition, and the calculation of subcellular localization data is particularly important. The possibility of an individual fruit fly representing a key protein can be measured by the odor concentration judgment function:

$$POS(i) = \eta \times OC(i) + (1-\eta) \times \sum_{j=1}^{q} DWD(a_j) \tag{23}$$

Suppose: $DWD(a_j)$ represents the subcellular localization score of the j-th protein in q key protein candidate sets, weight value η is used to adjust the importance degree of topological and biological features of the network to the key protein recognition results, and its value range is [0, 1]. If η = 1, then the key protein recognition results are only determined by the topological features of the network; if η = 1, then the key protein recognition results are only determined by the subcellular localization information in the biological features of the network.

Experimental results and analysis

It is known that, in the proposed key protein recognition algorithm, weight value η can realize the adjustment of the proportion of the importance of the topological and biological information of the network to the key protein recognition results. This paper analyzed the influence of the constant changing weight value η on the key protein recognition results, and the analysis results are shown in Table 1. According to the table, when the value range of weight value η is [0.5, 1], there's little difference in protein recognition results, after comprehensive consideration, its value was set to 0.5 in this paper.

Table 1. Influence of weight value η on recognition results

TOP / η	1%	5%	10%	15%	20%	25%
0	36	180	350	445	562	610
0.1	46	186	341	452	543	629
0.2	43	182	341	446	536	625
0.3	41	185	316	430	533	625
0.4	41	184	305	422	532	620
0.5	41	186	298	417	530	620
0.6	41	180	289	416	536	620
0.7	38	180	290	410	532	618
0.8	41	178	298	412	532	619
0.9	40	178	289	400	530	617
1	40	179	287	416	555	618

Table 2 shows the experimental results under different similarity thresholds of gene ontology annotations. Proteins with an odor concentration judgment value greater than threshold W in a round of iteration were taken as recognized proteins with strong interaction effect and were added into the seed set to perform the next round of iteration. According to the table, the experimental results of this method were good. The accuracy reached the highest 71.99% when

the similarity threshold of gene ontology annotations and threshold *W* were both 0.6. The recall rate reached the highest 72.42% when the similarity threshold of gene ontology annotations was 0.5 and the threshold *W* was 0.4. The F-score value reached the highest 70.88% when the similarity threshold of gene ontology annotations was 0.6 and the threshold *W* was 0.5. According to these results, when the similarity threshold of gene ontology annotations was set to 0.6, the protein network recognition results were better.

Table 2. Experimental results under different similarity values of gene ontology annotations

Threshold *W*	0.3			0.4		
Threshold of similarity	0.4	0.5	0.6	0.4	0.5	0.6
Accuracy	65.12	64.45	66.57	66.46	67.32	68.23
Recall rate	69.03	70.21	68.33	68.25	69.12	67.80
F-score	67.21	68.01	67.12	67.21	68.13	67.91
Threshold *W*	0.5			0.6		
Threshold of similarity	0.4	0.5	0.6	0.4	0.5	0.6
Accuracy	68.31	69.23	69.71	69.41	69.98	71.99
Recall rate	72.42	65.66	64.74	63.25	63.32	64.25
F-score	66.78	67.50	67.23	66.41	70.88	66.51

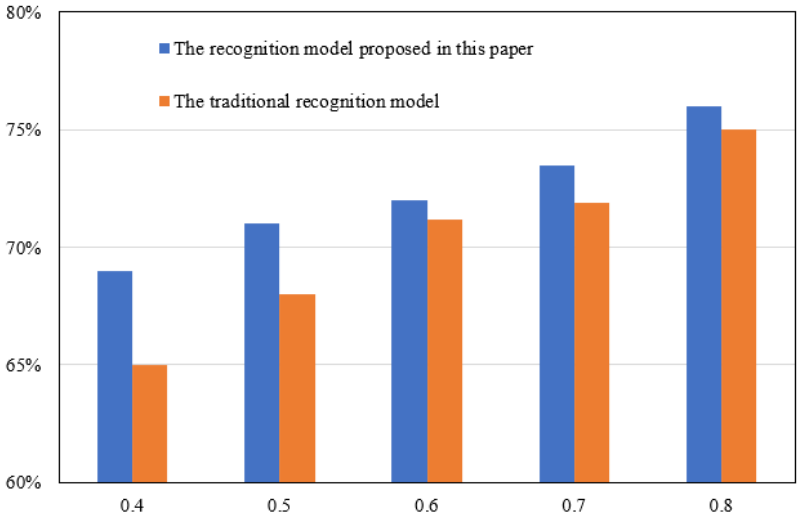
Table 3 shows the results of different iteration numbers when the similarity threshold of gene ontology annotations was set to 0.6, which further verified that the proposed method had achieved good protein recognition effect.

Table 3. Experimental results under different iteration numbers

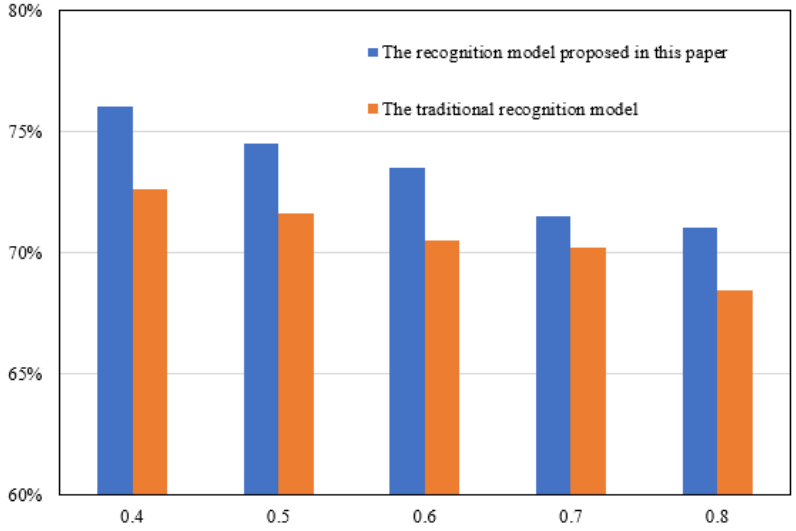
Threshold *W*	0.3			0.4		
Number of iterations	1	2	3	1	2	3
Accuracy	66.23	65.45	65.66	68.77	67.31	67.26
Recall rate	69.30	70.13	70.35	67.35	69.12	69.11
F-score	67.78	67.88	67.89	67.99	68.26	69.86
Threshold *W*	0.5			0.6		
Number of iterations	1	2	3	1	2	3
Accuracy	71.23	69.21	69.45	70.21	69.93	69.87
Recall rate	65.03	65.63	65.71	63.73	72.54	64.38
F-score	67.05	67.45	67.78	66.75	66.88	66.98

The highest accuracy 71.23% appeared in the first round of iteration when the *W* value was 0.5. The highest recall rate 72.54% appeared in the second round of iteration when the *W* value was 0.6. The highest F-score value 69.86% appeared in the third round of iteration when the *W* value was 0.4.

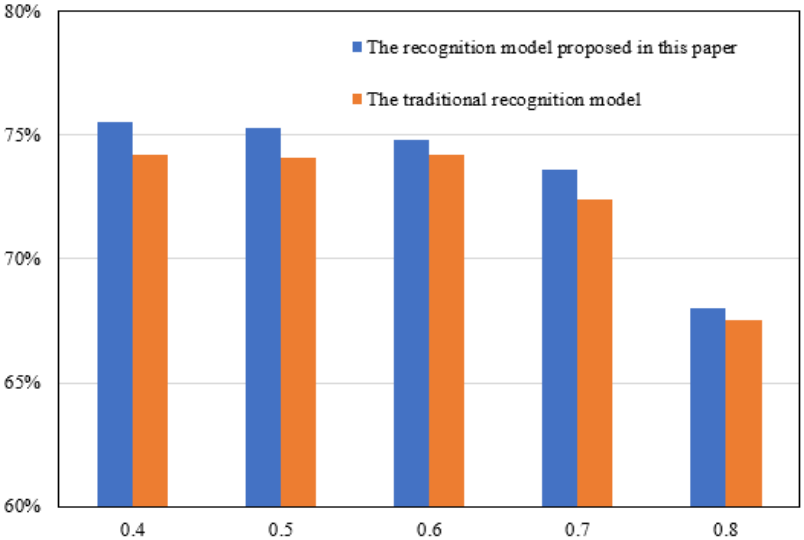
In Fig. 5, images a), b), c) respectively correspond to the experimental results of accuracy, recall rate, and F-score value of different recognition models, showing the comparison of the experimental results of the third iteration of the traditional weakly-supervised recognition model and the recognition model proposed in this paper. As can be seen, compared with the traditional recognition model, the protein recognition accuracy and F-score value of the proposed model had been improved significantly, while the recall rate showed a decline, wherein, the F-score value reached the highest in the third round of iteration when the *W* value was 0.4, which was 2.38% higher than the traditional weakly-supervised recognition model.



a)



b)



c)

Fig. 5 Comparison of experimental results of different recognition models

In order to further evaluate the proposed key protein recognition model, this paper plotted accuracy-recall curves to compare the proposed model with other models in terms of degree centrality, subgraph centrality, and local neighbor connectivity. As shown in Fig. 6, curve 1 represents the proposed model, curves 2, 3, and 4, respectively, correspond to recognition models based on local neighbor connectivity, subgraph centrality, and degree centrality. Obviously, the proposed model obtained the best effect, which had proved that the FOA proposed in this paper had a good effect in the recognition of key proteins in the high-order dynamic protein network.

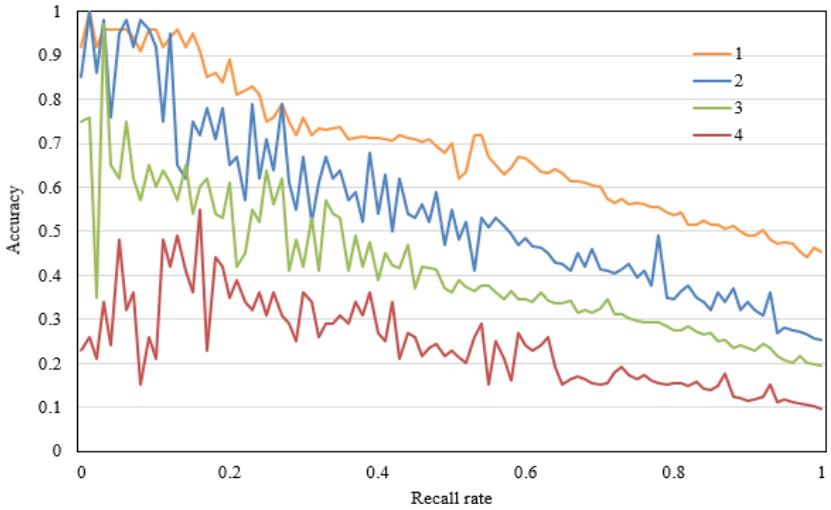


Fig. 6 Comparison of accuracy-recall curves of different recognition models

In order to further verify the recognition performance of the proposed key protein recognition model, this paper employed the jackknife method to compare the proposed model with other three models. In the jackknife method, a model with a bigger area under the curve has better recognition performance. As shown in Fig. 7, the area under the curve of the proposed model is the largest, which is much bigger than the other three models, and this can further verify the effectiveness of the proposed model in key protein recognition.

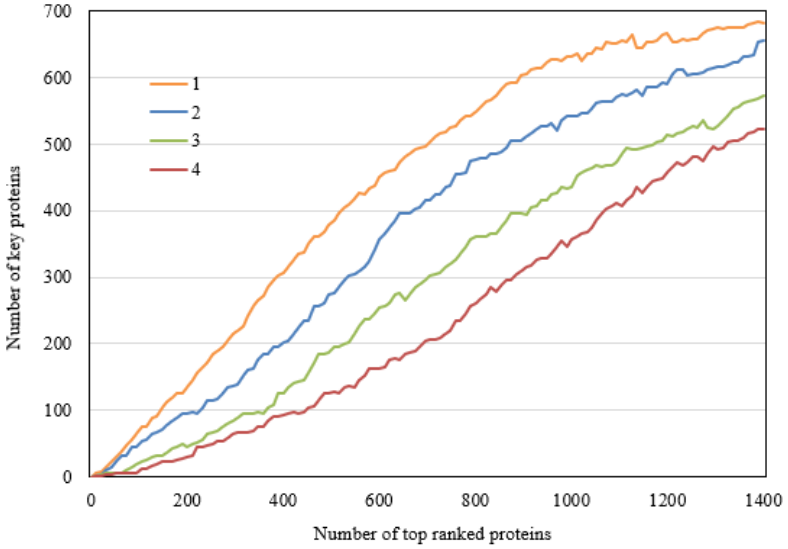


Fig. 7 Comparison of jackknife curves of different recognition models

Conclusion

This paper performed key protein recognition based on high-order dynamic complex protein network. At first, the paper elaborated on the method of feature selection of complex protein network and candidate set evaluation, and constructed a weighted protein network based on the obtained topological features of the complex protein network and the semantic similarity of protein gene ontology annotations. Then, the protein's active period was introduced into the constructed complex protein network to generate a high-order dynamic protein network model, and the proposed FOA completed effective recognition of key proteins in the high-order dynamic protein network. Combining with experiments, the experimental results obtained under conditions of different gene ontology annotation similarity and different iteration times were given, and the experimental results of accuracy, recall rate, and F-score of different recognition models were analyzed. At last, this paper plotted the accuracy-recall curves and jackknife curves of different recognition models, and the effectiveness of the proposed model had been verified.

References

1. Ceri S., A. Bernasconi, A. Canakoglu, A. Gulino, A. Kaitoua, M. Masseroli, P. Pinoli (2017). Overview of GeCo: A Project for Exploring and Integrating Signals from the Genome, Communications in Computer and Information Science, 822, 46-57.
2. Du L., J. Zhang, F. Liu, H. Wang, L. Guo, J. Han, (2021). Alzheimer's Disease Neuroimaging Initiative. Identifying Associations among Genomic, Proteomic and Imaging Biomarkers via Adaptive Sparse Multi-view Canonical Correlation Analysis, Medical Image Analysis, 70, 102003.
3. Exposito-Alonso M., H. G. Drost, H. A. Burbano, D. Weigel (2020). The Earth BioGenome Project: Opportunities and Challenges for Plant Genomics and Conservation, The Plant Journal, 102(2), 222-229.
4. Farnham P. J. (2012). Thematic Minireview Series on Results from the ENCODE Project: Integrative Global Analyses of Regulatory Regions in the Human Genome, Journal of Biological Chemistry, 287(37), 30885-30887.
5. Geerts M., A. Schnaufer, F. Van den Broeck (2021). rKOMICS: An R Package for Processing Mitochondrial Minicircle Assemblies in Population-scale Genome Projects, BMC Bioinformatics, 22(1), 1-14.
6. Grajeda C., L. Sanchez, I. Baggili, D. Clark, F. Breitinger (2018). Experience Constructing the Artifact Genome Project (AGP): Managing the Domain's Knowledge One Artifact at a Time, Digital Investigation, 26, S47-S58.
7. Hekmatnia E., H. Sajedi, A. Habib Agahi (2020). A Parallel Classification Framework for Protein Fold Recognition, Evolutionary Intelligence, 13(3), 525-535.
8. Huang H., J. Wang, J. Peng (2009). Recognition Essential Protein Based on Multi Parameters Combination, First International Workshop on Education Technology and Computer Science, 3, 861-866.
9. Jayaraj S., M. Gittelman (2018). Scientific Breakthroughs and Patent Scope: The Impact of the Human Genome Project on Early Stage Drug Patents, Proceedings of the 78th Annual Meeting of the Academy of Management, AOM 2018, https://doi.org/10.5465/AMBPP.2018.234.
10. Jiang J., M. Yan, D. Li, J. Li (2019). Genome Tagging Project: Tag Every Protein in Mice through 'Artificial Spermatids', National Science Review, 6(3), 394-396.
11. Khan A., A. Tyagi (2021). Considerations for Initiating a Wildlife Genomics Research Project in South and South-East Asia, Journal of the Indian Institute of Science, 101(2), 243-256.
12. Lei X., W. Zhang (2021). Logistic Regression Algorithm to Identify Candidate Disease

Genes Based on Reliable Protein-protein Interaction Network, Science China Information Sciences, 64(7), 1-3.
13. Nesbitt G., K. McKenna, V. Mays, A. Carpenter, K. Miller, M. Williams (2013). The Epilepsy Phenome/Genome Project (EPGP) Informatics Platform, International Journal of Medical Informatics, 82(4), 248-259.
14. Park S. Y., Z. Vaghchhipawala, B. Vasudevan, L. Y. Lee, Y. Shen, K. Singer, S. B. Gelvin (2015). Agrobacterium T-DNA Integration into the Plant Genome Can Occur without the Activity of Key Non-homologous End-joining Proteins, The Plant Journal, 81(6), 934-946.
15. Qu X., A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, K. A. Persson, (2015). The Electrolyte Genome Project: A Big Data Approach in Battery Materials Discovery, Computational Materials Science, 103, 56-67.
16. Saito K., N. Miura, M. Yamazaki, K. Tatsuguchi, M. Kurosawa, R. Kanda, I. Murakoshi (1994). Molecular Cloning and Expression of Key Enzymes for Biosynthesis of Cysteine and Related Secondary Non-protein Amino Acids, Schripsema J., R. Verpoorte (Eds.) Primary and Secondary Metabolism of Plants and Cell Cultures III, Springer, Dordrecht, 153-158.
17. Srivastava K., A. S. Fratzscher, B. Lan, W. A. Flegel (2021). Cataloguing Experimentally Confirmed 80.7 kb-long ACKR1 Haplotypes from the 1000 Genomes Project Database, BMC Bioinformatics, 22(1), 1-13.
18. Wahab Khattak F., Y. S. Alhwaiti, A. Ali, M. Faisal, M. H. Siddiqi (2021). Protein-protein Interaction Analysis through Network Topology (Oral Cancer), Journal of Healthcare Engineering, 2021, Article ID 6623904.
19. Wiedmann M. M., Y. S. Tan, Y. Wu, S. Aibara, W. Xu, H. F. Sore, D. R. Spring (2017). Development of Cell-permeable, Non-helical Constrained Peptides to Target a Key Protein-protein Interaction in Ovarian Cancer, Angewandte Chemie International Edition, 56(2), 524-529.
20. Yan K., J. Wen, J. X. Liu, Y. Xu, B. Liu (2020). Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(5), 2008-2016.
21. Zaki N., H. Singh, E. A. Mohamed (2021). Identifying Protein Complexes in Protein-protein Interaction Data Using Graph Convolutional Network, IEEE Access, 9, 123717-123726.
22. Zhao Y., P. Xie, H. Fan (2012). Genomic Profiling of MicroRNAs and Proteomics Reveals an Early Molecular Alteration Associated with Tumorigenesis Induced by MC-LR in Mice, Environmental Science & Technology, 46(1), 34-41.
23. Zhou T., K. H. Thung, M. Liu, D. Shen (2018). Brain-wide Genome-wide Association Study for Alzheimer's Disease via Joint Projection Learning and Sparse Regression Model, IEEE Transactions on Biomedical Engineering, 66(1), 165-175.

Xiaojing Wang
E-mail: 200800553@hhvc.edu.cn

Xiaojing Wang was graduated from Computer College, Inner Mongolia University, and now serves as a Chief of Integrated Management Division, Information Management Center, Hohhot Vocational College. Her research directions include computer science and technology and computer network. She chaired one provincial/ministerial project and participated in one municipal/departmental project.