Searching essential proteins in dynamic protein networks based on an improved H-index algorithm

Caiyan Dai (njucmdai@163.com)
Nanjing University of Chinese Medicine https://orcid.org/0000-0003-3562-3905
HE Ju
Nanjing University of Chinese Medicine
HU Kongfa
Nanjing University of Chinese Medicine
DING Youwei
Nanjing University of Chinese Medicine

Research article

Keywords: essential proteins; dynamic protein network; attenuation coefficient; improved H-index algorithm

Posted Date: November 6th, 2019

DOI: https://doi.org/10.21203/rs.2.16891/v1

Version of Record: A version of this preprint was published on June 17th, 2020. See the published version at https://doi.org/10.1186/s12911-020-01141-x.

Searching essential proteins in dynamic protein networks based on an improved H-index algorithm

DAI Caiyan[1], HE Ju[2], HU Kongfa[3] and DING Youwei[4]

[1,2,3,4] College of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine University, Nanjing 210000, China

Abstract

The essential proteins in protein networks play an important role in complex cellular functions and their evolution. Therefore, searching essential proteins in protein networks can help to explain the structure, function and dynamics of basic cellular networks. The existing dynamic protein network regards the protein situation at all times as the same, but in fact, the role of proteins varies at different times. In order to improve the accuracy of essential protein searching, an improved H-index algorithm based on attenuation coefficient method is proposed in this paper, which incorporates the neglected node information into consideration to improve the accuracy of essential protein searching. The experiments show that the essential proteins found on the basis of this model are more effective than other similar methods.

Key words: essential proteins; dynamic protein network; attenuation coefficient; improved H-index algorithm

1. Background

The essential proteins in protein networks play an important role in complex cellular functions and their evolution[1,2]. Therefore, searching the essential proteins in protein networks can help to explain the structure, function and dynamics of basic cellular networks.

In recent years, there have been some methods to mine essential nodes in complex networks from different perspectives. Wang et al. [3] proposed an effective method to identify vertices in dynamic networks using local detection and update strategies. This

[1] Correspondent author; Dai Caiyan, email: njucmdai@163.com

method detects the change vertices locally in the dynamic network and updates the influence measure of the change vertices locally, without calculating the influence of all vertices globally. Li et al. [4] proposed a new method for identifying essential proteins by combining protein complex information and PPI network topological features. By analyzing the relationship between protein complexes and essential proteins, it was found that proteins in multiple complexes were more likely to be essential than those in single complexes. Based on the statistical analysis of protein and protein complexes, Luo et al. [5] proposed a method for predicting essential proteins in protein-protein interaction networks based on local interaction density binding protein complexes. Hu et al. [6] proposed a new method, E-Burt method, which can be applied to weighted networks. The method fully considers the total connection strength, the number of connection edges and the distribution of the total connection strength on the connection edge in the local range. Wang et al. [7] used the iterative information of K-shell decomposition to distinguish the influence ability of nodes with the same k-shell. Many of the key factors to measure whether a node in complex networks is based on graph theory are to quantify the topological structure and attributes of each node, and to compare the centrality of each node through different centrality calculation methods, such as degree center, median center, proximity center and edge clustering coefficient center. Quantitative method finds the essential nodes in the network.

Searching essential proteins in dynamic protein networks can identify proteins that play the most important role in the evolution of proteins. When searching for key proteins, the above methods only consider the importance of nodes themselves to illustrate their centrality, ignoring some structural information on network graphs. So, the importance of the node itself and the rest of the structure information on the network can be combined together to examine the importance of the current node in the whole network.

2. Methods

There are essential proteins in the protein network, which are usually located at the center of the whole network. The appearance or destruction of these proteins is crucial to the impact of the whole protein network. Finding the essential proteins in dynamic protein networks accurately and quickly is helpful to understand various biological processes from a systematic point of view, and can be widely used to explore

the pathogenesis of diseases, predict and evaluate the corresponding treatment methods. At the same time, it can also find new drug targets and open up a new way for new drug research and development. Although there are some effective methods to find essential proteins in protein interactions, such as data mining, machine learning and artificial neural network, which have been applied to the research of essential protein search, it is still necessary to carry out in-depth research on effective algorithms to search essential proteins accurately and quickly.

(1) Time series on dynamic protein networks

When modeling dynamic protein networks, gene expression data and large-scale static protein networks are usually considered together. The gene expression arrays of M genes at T time points can be divided into T sets. Each set represents the state of the M genes at the same time point, and can be combined into a dynamic protein network based on time series, as shown in figure 1.

(2) Evolution of proteins

At different time points, there are different protein interaction networks. Figure 2 shows a simple protein evolution process.

The ultimate goal of the model is to facilitate subsequent research, such as identifying essential proteins that play a crucial role in the evolution of proteins, or predicting the linkages between proteins at the next moment. This requires recording the evolution of the protein itself. In the past, when recording the link relationship of protein network, the link relationship between proteins at different time points was directly recorded as 1, and the relationship without link was recorded as 0. This approach is inaccurate in time series, because over time, the role of historic protein data is declining, and the recently generated links between proteins play a larger role.

Dynamic protein networks represent the implementation of the whole evolutionary process over time. If the link relationship between proteins is considered to be the same in the evolutionary process over time, it will obviously affect a series of subsequent studies based on dynamic protein models. Therefore, it is worth exploring how to incorporate the protein-link relationship that changes with time and specific gravity into the scope of the model, so as to correctly identify key proteins in dynamic protein networks and predict the protein-link relationship at the next moment.

(3) The H-index method

H-index itself is a new method to evaluate academic achievements. H stands for

"high citation times". A scientist's *H*- index means that he has at most *H* papers cited at least *H* times respectively. The *H* index was originally used to accurately reflect a person's academic achievements. The higher a person's *H* index, the greater his academic influence. In the study of dynamic protein network, *H*-index can be used to find essential proteins, and formula (1) can be used to calculate the *H*-index value of nodes.

$$H-index(v_i)=H(d_{u1},d_{u2},...,d_{ud_i}) \quad uj \in neighbor(v_i) \tag{1}$$

Where d_i denotes the degree of node v_i, formula H (x_1, x_2,... x_n) returns the maximum value of y, at least y items come from x_1, x_2,... x_n are greater than or equal to y.

However, when using *H*-index, it only considers the importance of the node itself to illustrate its centrality, ignoring part of the structure information on the network, which will reduce the accuracy of node expansion ability.

For example, in figure 3, the centrality of node 2 is relatively large. However, considering the information of nodes in this graph, the expansion ability of node 3 is greater than that of node 2. The reason is that in the centrality of *H*-index, some information about node neighbors is ignored. For example, nodes with degree less than *y* are completely ignored, resulting in reduced specification accuracy of node expansion capability. Therefore, this method is not accurate in calculating the expansion capacity of nodes.

Therefore, it is necessary to improve the existing *H*-index algorithm on the basis of the established dynamic protein network model in order to quickly and accurately find the key proteins in the dynamic protein network by combining the information of the node itself and some of the structural information neglected in previous algorithms.

（4）Monotonicity

The ability to distinguish nodes with different scalability and nodes with uniform distribution at different levels is one of the criteria for evaluating the ranking methods of influential nodes in social networks [8]. Monotonicity is used to test the recognition ability of this method for nodes with different extensibility. The formula (2) is used to calculate the m value of ranking table R. In this equation, n is the number of column groups on list R and the number of nodes on column group R. The value of M is always

a number in the range [0,1]. Large numbers indicate that nodes have high recognition ability.

$$M(R) = (1 - \frac{\sum_{r \in R} n_r * (n_r - 1)}{n * (n-1)})^2 \qquad (2)$$

Based on the established protein network based on time-attenuation model, the essential protein recognition methods are studied. The methods to be adopted are as follows.

(1) Constructing the whole network of protein evolution process based on attenuation coefficient

In this network, the weights of each side at the corresponding time should be added together to get the final weights. The formulas are as follows:

In the process of construction, the same edge appears at different times. At the current time, the corresponding weight calculation is different. The earlier the edge appears, its role in the whole process of protein evolution will change over time. The calculation method of weight corresponding to edges at each time is as follows:

For each edge (u, v) in the protein network at time t, its weight D (u, v, t) is a variable with time t, which is defined as:

$$D((u,v),t) = \begin{cases} \delta((u,v),0) & t = 0 \\ D((u,v),t-1) * \lambda + \delta(I,t) & otherwise \end{cases} \qquad (3)$$

in which

$$\delta((u,v),t) = \begin{cases} 1 & a(t)\ include\ edge\ (u,v) \\ 0 & otherwise \end{cases} \qquad (4)$$

a (t) is a set of all the edges appearing at time t, which is a constant, called attenuation coefficient.

The weight of each vertex neighbor is calculated by the weight of the edge. If the vertex v is a neighbor of u, the weight w (u, v) of v for u is defined as:

$$w(u,v) = \frac{D(u,v,t)}{\sum_{x \in N(u)} D(u,x,t)} \qquad (5)$$

It can be seen from the above definition that $\sum_{v \in N(u)} w(u,v,t)=1$ and w (u, v) and w (v, u) are not necessarily equal.

In this way, the weight of each edge in the protein network at different time is different.

The construction process of dynamic protein network based on attenuation coefficient is shown in figure 4.

Considering the changes of historic proteins with time in the process of protein evolution, the protein network model is more objective and conforms to the process of biological evolution.

(2) Calculating the cumulative centrality of node neighborhood

Although H-index measures try to determine the centrality of nodes based on the importance of adjacent nodes, some information about adjacent nodes is still ignored. The centrality of a node can be standardized by using all the information of its adjacent nodes. For this purpose, the cumulative function in definition 1 is used.

Definition 1. The cumulative function $c_k(v_i)$ is defined as the number of nodes whose v_i neighbors are moderately larger than or equal to k, expressed as follows:

$$c_k(v_i) = \left| \{ v_j \mid v_j \in N_i \ and \ d_j \geq k \} \right| \tag{6}$$

The H-index function is improved to the cumulative function defined in equation (7).

$$H - index(v_i) = \overset{argmax}{k}(c_k(v_i) \geq k) \tag{7}$$

The function takes the maximum value of k, where the degree of k neighbors is greater than or equal to k when other cumulative values are ignored by H-index function. Therefore, in the proposed metric, the cumulative value of neighbor nodes is determined for all different k, and it is used to standardize the centrality of nodes.

In the traditional H-index, the weight of each neighbor's link is considered as 1. In our problem, the weight of each neighbor's link $w(u,v)$ is different. Therefore, when extending the traditional H-index, the extended cumulative function $c_k(v_i)$is defined as follows:

Definition 1'. Extended cumulative function $c_k(v_i)$is defined as:

$$c_k(v_i) = \sum_{v_j \in N_k(v_i)} w(v_i, v_j) \tag{8}$$

Here $N_k(v_i)$ is the set of neighbor vertices whose degree is k.

Definition 2 cumulative function vector $S(v_i)$ contains the cumulative values of adjacent nodes of node v_i at different degrees. The calculation method is shown in formula (9).

$$s(v_i) = \{c_1(v_i), c_2(v_i), ..., c_h(v_i)\} \tag{9}$$

Among them, h is the largest degree on the graph, and $h = \max_{j=1,...,n}\{d_j\}$.

In order to reduce the time complexity of calculating vector $S(v_i)$ elements, the recursive equation (10) can be used.

$$s_k(v_i) = \begin{cases} \sum_{v_j \in N(v_i)} w(v_i, v_j) & if \quad k = 1 \\ s_{k-1}(v_i) - \sum_{v_j \in N_{k-1}(v_i)} w(v_i, v_j) & if \quad k > 1 \end{cases} \tag{10}$$

Here, $S_k(v_i)$ is the k-th index value of vector $S(v_i)$, and $N_k(v_i)$ is the set of neighbor vertices whose degree is k.

Given the cumulative function vector of node v_i, its cumulative centrality is expressed as equation (11).

$$CMC(v_i) = \sum_{k=1}^{h} p^{1+k*\frac{p}{r}} * s_k(v_i) \tag{11}$$

In this formula, p and r are two adjustable parameters, and the value of p is between [0,1]. Because there is a larger cumulative value in the lower degree in the higher degree, and in the higher order cumulative value of many nodes, equation (11) uses a parameter $p^{1+k*\frac{p}{r}}$ to multiply the lower order cumulative value by a larger number. This ensures that the lower-order cumulative value is more effective, and has stronger expansion and recognition ability in the regulation of node centrality.

(3) The extended H-exponential centrality EHC (v) of the node is determined according to the cumulative centrality of its neighborhood.

Formula (12) can be used to determine the extended H-index centrality EHC (v) of nodes by iteration.

$$\begin{aligned} EHC^{(0)}(v) &= CMC(v) \\ EHC^{(t+1)}(v) &= \sum_{u \in N(v)} w(v,u).EHC^{(t)}(v) \end{aligned} . \tag{12}$$

(4) Calculate the centrality of all nodes and arrange them in order. N nodes with larger centrality are the essential nodes.

3. Results

3.1 Experimental data

The data used in the experiment are as follows:

(1) Gene expression data GSE3431 [9], the corresponding matrix has 6470 lines, each line represents the corresponding expression data of different genes;

(2) Yeast protein network in DIP [10], which includes 5093 proteins and 24743 edges, is shown in figure 5.

(3) 1285 essential proteins obtained from the datasets MIPS [11] , SGD [12] ,DEG [13] and SGDP [14].

3.2 Experimental results

3.2.1 Experiments of parameter selection

Firstly, the attenuation coefficient is tested. The dynamic protein network was divided into 36 moments, and the attenuation coefficients were compared with different values. The proposed algorithm is abbreviated as IH-index.

The SIR extension model[8] is used to evaluate the accuracy of this method in determining the node expansion capability and sorting the nodes. For this reason, the diffusion process is simulated by SIR, and the real ranking table σ is generated. In the SIR process, each node can be in one of three states: susceptibility (S), infection (I) and recovery (Re). After the necessary changes in the state of all nodes, the node of state Re is considered as the extension capability of node v_i. Through repeated processing, the scalability of each node is calculated, and the ranking table σ is obtained.

After calculating the table σ, the sorting table R can be also generated by different methods. The higher the correlation between the two ranking tables, the higher the accuracy of the corresponding methods in the specification of node expansion capability. For this reason, the Kendall correlation coefficient τ $(0 \leq \tau \leq 1)$ is adopted.

$$\tau(\sigma, R) = \frac{n_c - n_d}{n(n-1)/2} \tag{13}$$

Among them, n_c and n_d denote the number of consistent and inconsistent pairs of nodes in the two sorting tables, and n denotes the size of the sorting vector. The larger the Kendall correlation coefficient τ values, the closer the relationship between the two tables σ and R, the more accurate the proposed algorithm for calculating essential degree of the dynamic proteins.

From the above two figures, it can be found that the number of essential proteins found and the accuracy of essential proteins are changing with the change of attenuation coefficient. By synthesizing the two experimental results, we can find that the attenuation coefficient ranged from 0.9 to 0.95, and the number of essential proteins

found was the best. Therefore, the attenuation coefficient was set to 0.92.

Next, the effect of parameters p and r on the results of the search algorithm is detected on yeast protein dataset by Kendall coefficient.

From table 1, it can be found that the value of Kendall correlation coefficient changes slightly with the change of parameters p and r. When $p = 0.9$ and $r = 100$, the maximum value of τ can be obtained. So, the following experiments are carried out in the case of $p = 0.9$ and $r = 100$.

3.2.2 Experimental results of dynamic protein network models based on attenuation coefficients for different algorithms

In order to verify the performance of dynamic protein network based on attenuation coefficient, different algorithms can be used to find essential proteins in the established network, and the results are compared.

The four essential node search methods: Cnc+ [14], H-index [15], IGC [16] and IH-Index are run on the constructed attenuation coefficient-based protein network.

(1) The monotonicity values

Next, we verify the monotony of different algorithms based on dynamic protein network to find essential proteins. The results are shown in figure 8.

From Figure 8, it can be found that the monotonicity value of IH-index algorithm is higher than other algorithms, close to 1, which shows that this algorithm has a stronger ability to recognize essential proteins.

(2) The correctness

Figure 9 shows the Kendall coefficients of two sorting tables corresponding to different algorithms. It can be seen from the figure that the accuracy of H-index algorithm in finding essential proteins is slightly higher than that of Cnc + and IGC, and significantly higher than that of H-index.

4. Discussion

The next step is to consider that the current situation of an edge plays a very small role in the network when the weight of the edge decreases with time and reaches a minimum threshold. It can be directly subtracted to save time and space. So, the minimum threshold is set for the weight of edges. The next step is also plan to apply this algorithm to the study of dynamic protein sequence data.

5. Conclusions

In order to consider the influence of historical data on current data during protein evolution, a dynamic protein network model based on attenuation coefficient is proposed. In this model, instead of simply generalizing the appearance or absence of proteins at each time, dynamic protein network modeling method based on attenuation coefficient is used to record the changes of proteins in the process of biological evolution according to their corresponding occurrences. The traditional key node search method, H-index algorithm, which neglects neighbor attributes, is improved. The cumulative function is used to take into account the different degree of neighbor attributes of nodes in the network, which makes the essential proteins search more accurate. In order to verify the validity, different key node search methods are used in dynamic protein network. The experimental results show that the model established by IH-index method in this paper is more convenient to find essential proteins accurately.

References

[1] Wang S, Cuomo S, Mei S, Cheng W, Xu N. Efficient method for identifying influential vertices in dynamic networks using the strategy of local detection and updating. Future Generation Computer Systems. 2019, 91: 10-24.

[2] Shanping Qiao, Baoqiang Yan, Jing Li. Ensemble learning for protein multiplex subcellular localization prediction based on weighted KNN with different features. Applied Intelligence. 2018,48(7): 1813–1824.

[3] Li M, Lu Y, Xiang N, Pan W. Identification of Essential Proteins by Using Complexes and Interaction Network. Bioinformatics Research and Applications. 2014,255-265.

[4] Luo J, Qi Y. Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes. PLoS One. 2015, 10(6):e0131418.

[5] Hu P， Mei T. Ranking Influential Nodes in Complex Networks with Structural Holes. Physica A: Statistical Mechanics and Its Application. 2018,490:624-631.

[6] Wang Z, Zhao Y, Xia J, Du C. Fast ranking influential nodes in complex networks using a k-shell iteration factor. Physica A: Statistical Mechanics and its Applications. 2016,461:171-181.

[7]Ahmad Z, Amir S. EHC: Extend H-index centrality measure for identification of users' spreading influence in complex networks. Physica A: Statistical Mechanics and its Applications. 2019:1-13.

[8] J. Bae and S. Kim. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Physica A: Statistical Mechanics and its Applications. 2014,395:549-559.

[9]https://www.ncbi.nlm.nih.gov

[10]http://dip.deo-mbi.ucla.edu/dip/Stat.cgi

[11]http://mips.helmholtz-muenchen.de/proj/ppi

[12] https://www.yeastgenome.org

[13]Zhang R,Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes[J]. Nucleic Acids Research. 2009,7:D455-D458.

[14]http://www-sequence. stanford. edu/ group/ yeast _ deletion _project / deletions3.html

[15] L. Lü, T. Zhou, Q.M. Zhang, and H. E. Stanley. The H-index of a network node and its relation to degree and coreness, Nature communications, 2016, 7:10168.

[16] J. Wang, C. Li, and C. Xia. Improved centrality indicators to characterize the nodal spreading capability in complex networks, Applied Mathematics and Computation. 2018, 334:388-400.

Figures legends :

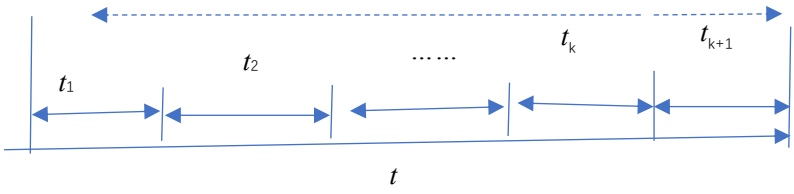


Figure 1 Time series diagram of dynamic protein network

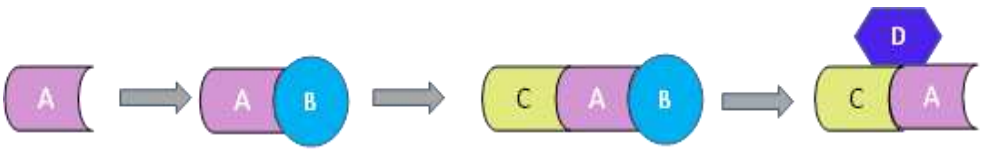


Figure 2 Simple Protein Evolution

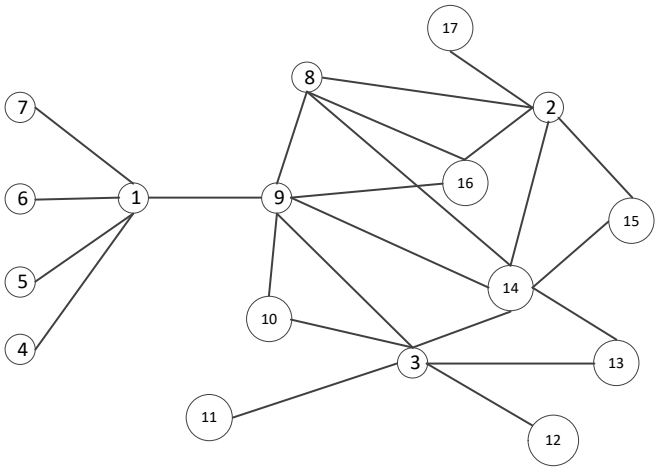
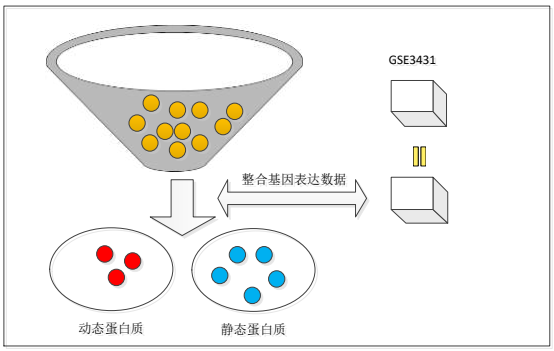
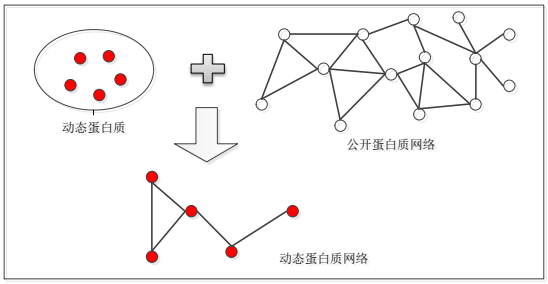


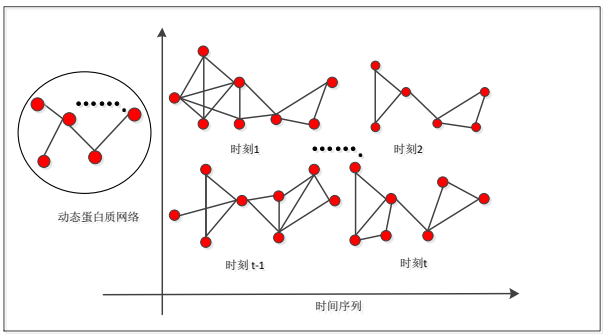
Figure 3 The example of node centrality



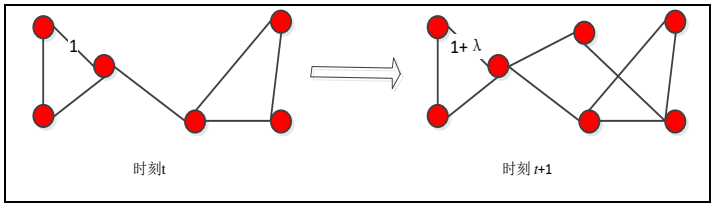
a. Integrating gene expression data and extracting dynamic proteins



b. Combining open protein network and dynamic protein to form dynamic protein network



c. Constructing dynamic protein networks at different times



d. The weight of the same edge at adjacent time

Figure 4 Construction of dynamic protein network based on attenuation coefficient

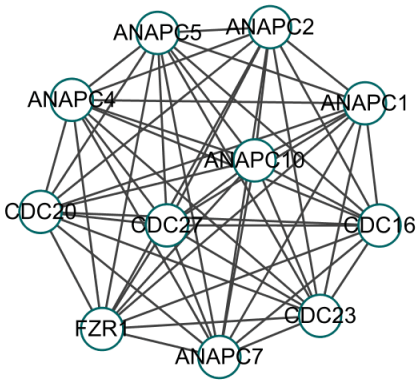


Figure 5 Part of yeast protein network

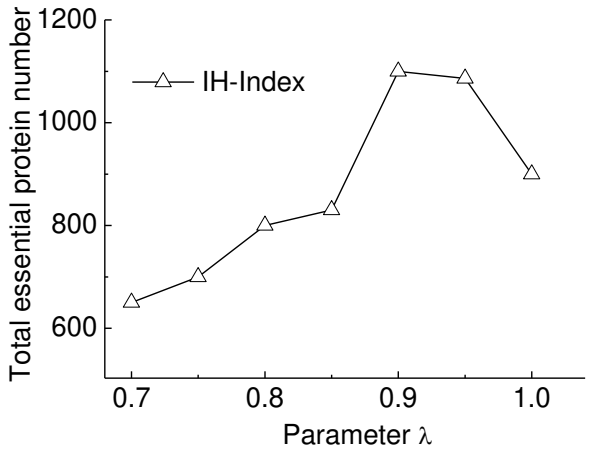


Figure 6 The effect of attenuation coefficient on the essential protein number

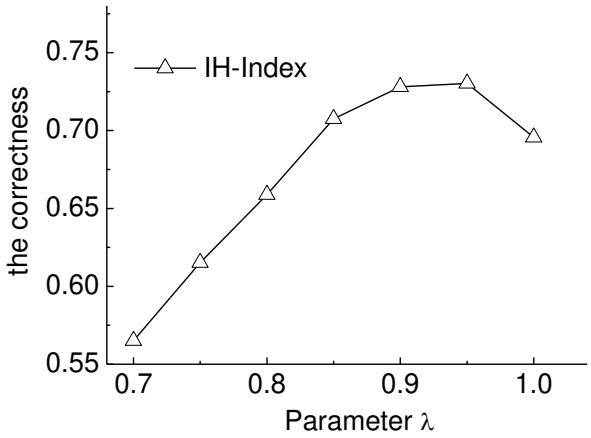


Figure 7 The correctness changes with the attenuation coefficient changes

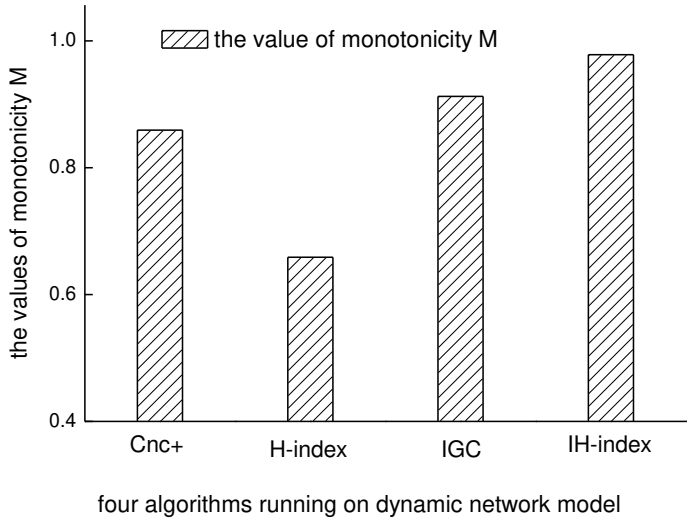


Figure 8 Monotonicity values of different essential protein search algorithms

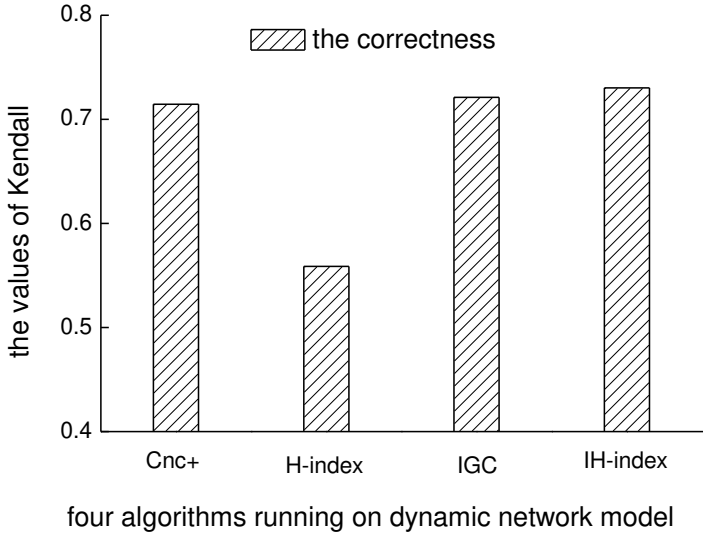


Figure 9 Correctness of different essential protein search algorithms

Table 1 Kendall coefficient values corresponding to different parameters p and r

p \ r	1	10	100	1000	10000
0.5	0.6784	0.7234	0.7312	0.7313	0.7320
0.6	0.6825	0.7242	0.7321	0.7321	0.7321
0.7	0.6856	0.7268	0.7315	0.7327	0.7322
0.8	0.6921	0.7285	0.7304	0.7307	0.7301
0.9	0.7172	0.7294	0.7335	0.7326	0.7303
1.0	0.7281	0.7291	0.7301	0.7292	0.7293

Figures

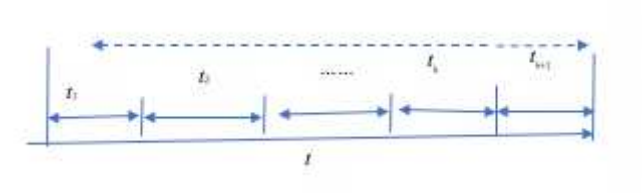


Figure 1

Time series diagram of dynamic protein network

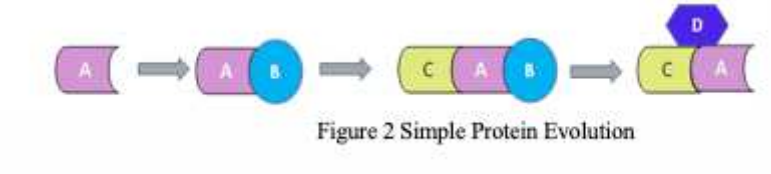


Figure 2 Simple Protein Evolution

Figure 2

Simple Protein Evolution

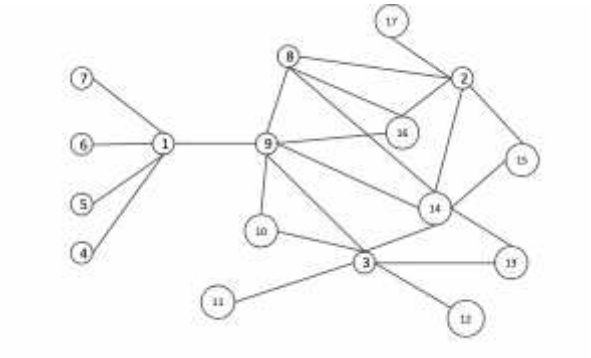
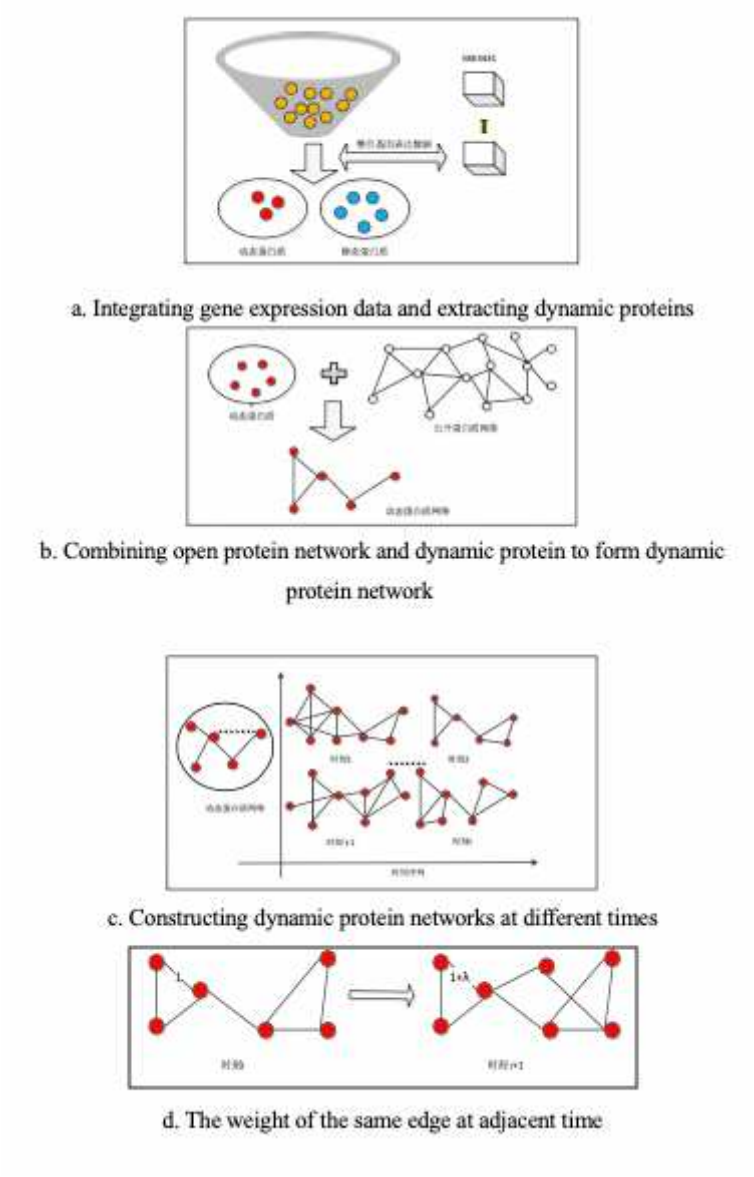


Figure 3

The example of node centrality



a. Integrating gene expression data and extracting dynamic proteins

b. Combining open protein network and dynamic protein to form dynamic protein network

c. Constructing dynamic protein networks at different times

d. The weight of the same edge at adjacent time

Figure 4

Construction of dynamic protein network based on attenuation coefficient

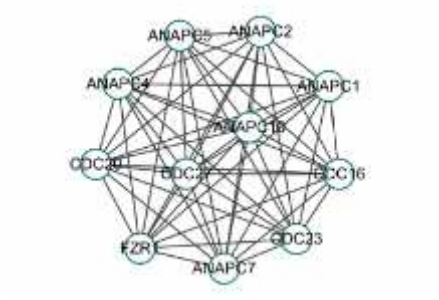


Figure 5

Part of yeast protein network

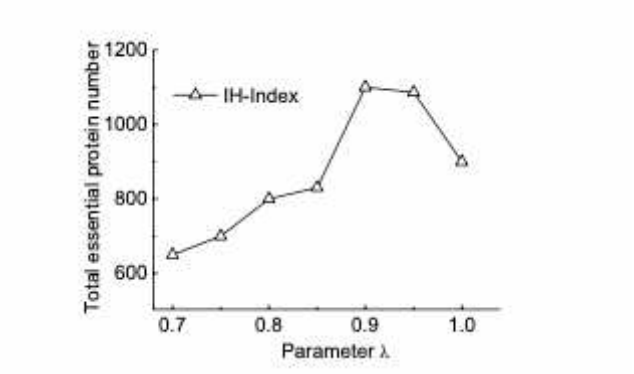


Figure 6

The effect of attenuation coefficient on the essential protein number

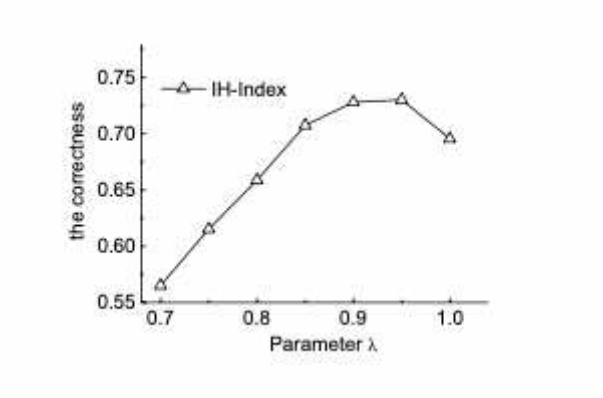


Figure 7

The correctness changes with the attenuation coefficient changes

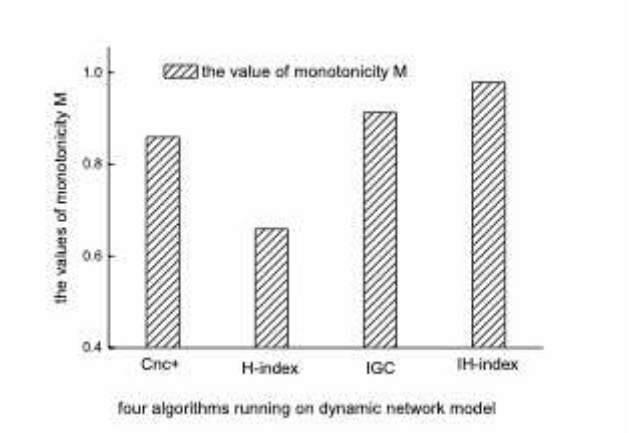


Figure 8

Monotonicity values of different essential protein search algorithms

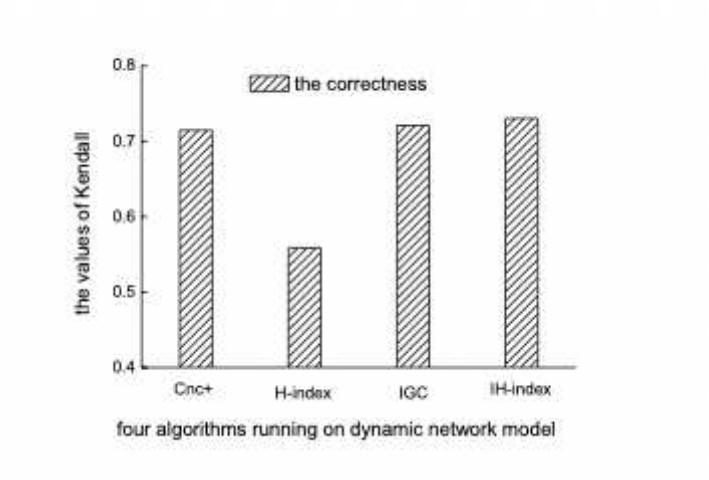


Figure 9

Correctness of different essential protein search algorithms