SYNOPSIS
Final Year Project
(CS 794)

IDENTIFICATION OF ESSENTIAL PROTEINS BASED ON PROTEIN-PROTEIN INTERACTION (PPI) NETWORK BY LBCC METHOD

**Ananya Chakraborty (06) , Anannya Bhattacharjee (05) ,
Alokananda Ghosh (04) , Dyuti Giri (17)**

WHAT IS PROTEIN?

Proteins are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues. Proteins perform a vast array of functions within organisms, including catalysing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells and organisms, and transporting molecules from one location to another. Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in protein folding into a specific three-dimensional structure that determines its activity.

PROTEIN-PROTEIN INTERACTION NETWORK (PPIN)

Protein-protein interactions (PPIs) are essential to almost every process in a cell, so understanding PPIs is crucial for understanding cell physiology in normal and disease states. It is also essential in drug development, since drugs can affect PPIs. Protein-protein interaction networks (PPIN) are mathematical representations of the physical contacts between proteins in the cell. These contacts:

- are specific;
- occur between defined binding regions in the proteins; and
- have a particular biological meaning (i.e., they serve a specific function).

WHY IS PPI NETWORK IMPORTANT?

Essential proteins are indispensable to the viability and reproduction of an organism. The identification of essential proteins is necessary not only for understanding the molecular mechanisms of cellular life but also for disease diagnosis, medical treatments and drug design. Many computational methods have been proposed for discovering essential proteins, but the precision of the prediction of essential proteins remains to be improved. Hence, PPIN is important in research for identification of essential proteins.

Knowledge of PPIs can be used to:

- assign putative roles to uncharacterised proteins;
- add fine-grained detail about the steps within a signalling pathway; or
- characterise the relationships between proteins that form multi-molecular complexes such as the proteasome.

PAPER STUDY

“A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes”

Chao Qin, Yongqi Sun*, Yadong Dong

OVERLAY STRUCTURE

Essential proteins are indispensable to the viability or reproduction of an organism and play a decisive role in cellular life. Deletion of a single essential protein is sufficient for causing lethality or infertility. Compared to non-essential proteins, essential proteins are more likely to be conserved in biological evolution. Essential proteins provide insights into the molecular mechanisms of an organism at the system level, with significant implications for drug design and disease study. For example, in drug development, essential proteins are excellent targets for potential new drugs and vaccines to treat and prevent diseases and for improved diagnostic tools more reliably to detect infections. The identification of essential proteins is necessary not only for understanding the molecular mechanisms of cellular life but also for disease diagnosis, medical treatments and drug design. Many computational methods have been proposed for discovering essential proteins, but the precision of the prediction of essential proteins remains to be improved. In this paper, we propose a new method, LBCC, which is based on the combination of local density, betweenness centrality (BC) and in-degree centrality of complex (IDC). First, we introduce the common centrality measures; second, we propose the densities Den1(v) and Den2(v) of a node v to describe its local properties in the network; and finally, the combined strategy of Den1, Den2, BC and IDC is developed to improve the prediction precision. The experimental results demonstrate that LBCC outperforms traditional topological measures for predicting essential proteins, including degree centrality (DC), BC, subgraph centrality (SC), eigenvector centrality (EC), network centrality (NC), and the local average connectivity-based method (LAC). LBCC also improves the prediction precision by approximately 10 percent on the YMIPS and YMBD datasets compared to the most recently developed method, LIDC.

DATASET USED

To evaluate the performance of the LBCC method, *Saccharomyces cerevisiae* is used as the experimental material because relatively reliable and complete PPI data are available for this organism. The PPI network data are from the MIPS database (Mammalian Protein-Protein Interaction Database), the DIP database, and other datasets from the website of the Mark Gerstein Lab (gersteinlab.org). Four different datasets are selected. The first dataset, YMIPS; the second and third datasets from the Mark Gerstein Lab were marked YMBD and YHQ, respectively; and the fourth dataset, was marked YDIP. YMIPS included 4546 proteins and 12319 interactions, and its average degree was approximately 5.42. YMBD, includes 2559 proteins and 11835 interactions, and its average degree was approximately 9.25. YHQ includes 4743 proteins and 23294 interactions in total. The average degree of YHQ was approximately 9.82. YDIP included 5093 proteins and 24743 interactions, and its average degree was approximately 9.72.

Information on the four PPI datasets: YIMP,YMBD, YHQ, and YDIP.

Dataset	Proteins	Interactions	Average degree	Essential proteins
YMIPS	4546	12319	5.42	1016
YMBD	2559	11835	9.25	783
YHQ	4743	23294	9.82	1108
YDIP	5093	24743	9.72	1167

METHODOLOGY

Notation

For an undirected simple graph G(V, E) with a set of nodes V and a set of edges E, a node v ϵ V denotes a protein and an edge e(u, v) ϵ E denotes an interaction between two proteins u and v. Nv denotes the set of nodes containing all the neighbors of node v, and |Nv| denotes the number of nodes in Nv. Let G[S] denote the subgraph of G induced by the node set S.

Centrality measures

Many researchers have found that it is significative to predict essential proteins by centrality measures. A PPI network is always represented as an undirected simple graph G(V, E). Here, we will introduce six classical centrality measures based on network topological properties.

- *Degree centrality*(DC). The degree centrality of a node v is the number of its neighbor nodes,

$$\mathrm{DC(v) = deg(v)},$$

where deg(v) is the number of its neighbor nodes.

- *Betweenness centrality*(BC). The betweenness centrality of a node v is denoted as the average fraction of the shortest paths passing through the node v,

$$\mathrm{BC(v)} = \sum_{s \neq t \neq v \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σst is the number of shortest paths between s and t, and σst(v) is the number of such paths passing through v.

- *Subgraph centrality*(SC). The subgraph centrality of a node v accounts for the participation of v in all subgraphs of the network,

$$\mathrm{SC(v)} = \sum_{k=0}^{\infty} \frac{\mu_k(v)}{k!}$$

where μk(v) is the number of subgraphs from node v to node v with length k.

- *Eigenvector centrality*(EC). The eigenvector centrality of a node v is the value of the vth component of the principal eigenvector of A,

$$\mathrm{EC} = \alpha_{max}(v)$$

where αmax represents the eigenvector that corresponds to the largest eigenvalue of the adjacency matrix A and αmax(v) is the vth component of αmax.

- *Local average connectivity centrality*(LAC). The local average connectivity centrality of a node v is denoted as the local connectivity of its neighbors,

$$\text{LAC(v)} = \frac{\sum_{u \epsilon N_V} deg^{cv}(u)}{|N_V|}$$

where C_v is the subgraph G[Nv] and deg^{C_v} (u) is the number of its neighbors in Cv for a node u ϵ Nv.

- *In-degree centrality of complex*(IDC).The in-degree centrality of complex of a node v is denoted as

$$\text{IDC(v)} = \sum_{i\,\epsilon ComplexSet(v)} IN - Degree(v)$$

where ComplexSet(v) represents the set of protein complexes including protein v and INDegree(v)i is represented as the value of DC(v) for the ith protein complex belonging to ComplexSet(v).

<u>Local properties of nodes in a PPI network</u>

There are many local properties of nodes in a PPI network, such as the degree centrality (DC) and *local clustering coefficient,* which is defined as

$$\text{LCC(v)} = \frac{2(|E(H)|-|Nv\,|)}{|Nv|(|Nv|-1)} :$$

In this section, we propose two types of local properties of nodes in a PPI network, Den1(v) and Den2(v), which are defined as follows.
Den1(v). For a node v, let H denote the subgraph of G[Nv [{v}]; then, we define

$$Den_1 = \frac{2|E(H)\,|}{|V(H)|(|V(H)|-1)}$$

which is the proportion of the number of the edges to the number of all possible edges of H. Den1(v) is somewhat different from LCC(v), and their relationship is

$$Den_1 = \frac{(|N_V|-1)LCC(v)+2}{(|N_V|+1)}$$

And Den2(v),

$$Den_2 = \frac{2|E(H)\,|}{|V(H)|(|V(H)|-1)}$$

Hence, Den2(v) is the density of the subgraph induced by v and the set of nodes for which the distance to v is 1 or 2. The local properties Den1(v) and Den2(v) are important for aiding in locating essential proteins.

New centrality measure: LBCC

In this section, we propose a new method, LBCC, by combining Den1, Den2, BC and IDC. The following basic concepts underlie LBCC:

1. essential proteins tend to form highly connected clusters;
2. essential proteins gather in protein complexes; and
3. both local and global properties are important for aiding in locating essential proteins.

Therefore, for a node v of the network, we use IDC(v) to represent its information on protein complexes and BC(v) to represent its global properties. For the contribution of local properties and highly connected clusters, we use Den1(v) and Den2(v). Because the value ranges of these measures differ, we apply a log transformation to normalize the data. Now, we can describe our new measurement LBCC for evaluating the essentiality of a node v,

LBCC(v) = a*logDen1(v) + b*logDen2(v) + c*logIDC(v) + d*logBC(v);

where a, b, c, and d are scaling parameters that range from 0 to 10 and represent the importance of the corresponding item used in the LBCC calculation. We set IDC(v) = 0.001 if a protein v does not appear in any protein complex.

In the YMIPS dataset, the measurement LBCC has the best performance when a, b, c and d are set to 1, 4, 3 and 1, respectively. IDC and BC are more important than Den1 and Den2 when calculating LBCC.

EXPLANATION

There are two types of methods for predicting essential proteins. One is experimental procedures, such as RNA interference, single gene knockouts, and conditional knockouts.

However, these experimental procedures require considerable time and resources, even for well-studied organisms, and they are not always practical. The other type of method is bioinformatics computational approaches that take advantage of the abundance of experimental data available for protein interaction networks, such as degree centrality (DC), betweenness centrality (BC), subgraph centrality (SC), eigenvector centrality (EC), network centrality (NC), and the local average connectivity-based method (LAC). Obviously, the latter is faster and less expensive than the former. In 2015, Luo and Qi proposed a method named LIDC for discovering essential proteins based on the local interaction density and protein complexes. The experimental results obtained with the YMIPS dataset demonstrated that the performance of LIDC was superior to that of nine reference methods (i.e., DC, BC, NC, LID, PeC, CoEWC, WDC, ION, and UC). However, methods based on bioinformatics computational approaches are sensitive to the local or global topological properties of the network, and the prediction precision for identifying essential proteins requires further improvement. In this paper, we first introduce the densities Den1(v) and Den2(v) of a node v to describe its local properties in the network. Then, a novel method called LBCC is proposed, which is combined with Den1, Den2, BC, and IDC, where the local and global properties of the node are measured by Den1 and Den2 and by BC, respectively, and the information of the protein complex is measured by IDC, which was first introduced in. This combination of features has not previously been considered for this problem. We performed several experiments on different PPI (protein-protein interaction) networks of Saccharomyces cerevisiae, YMIPS, YMBD, YHQ and YDIP, which will be described in the Experimental data section. The experimental results demonstrate that our LBCC method provides superior prediction performance compared to centrality measures, including DC, BC, SC, EC, NC, and LAC. In particular, compared to the most recent method, LIDC, which is a more effective method for predicting essential proteins, LBCC improves the prediction precision by at least 10 percent on the YMIPS and YMBD datasets.

RESULT AND DISCUSSION

Evaluation methods

In general, several statistical measures, such as sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure (F), and accuracy (ACC), are used to determine how effectively the essential proteins are identified by different methods including proposed method LBCC. First, we provide four statistical terms:

- *True positives*(TP). The essential proteins that are correctly selected as essential.
- *False positives*(FP). The nonessential proteins that are incorrectly selected as essential.
- *True negatives*(TN). The nonessential proteins that are correctly selected as nonessential.
- *False negatives*(FN). The essential proteins that are incorrectly selected as nonessential.

Next, we provide the definitions of six statistical measures:

1. *Sensitivity*. Sensitivity is the ratio of the proteins that are correctly selected as essential to the total number of essential proteins,

$$\mathrm{SN} = \frac{TP}{TP+FN}$$

2. *Specificity*. Specificity is the ratio of the nonessential proteins that are correctly selected as nonessential to the total number of nonessential proteins,

$$\mathrm{SP} = \frac{TN}{TN+FP}$$

3. *Positive predictive value*. Positive predictive value refers to the ratio of the proteins that are correctly selected as essential,

$$\mathrm{PPV} = \frac{TP}{TP+FP}$$

4. *Negative predictive value*. Negative predictive value refers to the ratio of the proteins that are correctly selected as nonessential,

$$\mathrm{NPV} = \frac{TN}{TN+FN}$$

5. *F-measure*. F-measure refers to the harmonic mean of SN and PPV,

$$\mathrm{SN} = \frac{2*SN*PPV}{SN+PPV}$$

6. Accuracy. Accuracy refers to the ratio of the proteins that are correctly selected as essential and nonessential in all the results,

$$\mathrm{ACC} = \frac{TP+TN}{P+N}$$

in which P represents the number of essential proteins and N represents the number of nonessential proteins

Comparison with other prediction measures

To evaluate the performance of LBCC, we compared LBCC and other prediction measures using the four datasets described in the Experimental data section. The compared prediction measures included LIDC, DC, BC, SC, EC, NC, and LAC. The algorithm for LIDC was implemented and the other algorithms were implemented using CytoNCA, a plugin of Cytoscape for centrality analysis of PPI networks. First, proteins are ranked in descending order based on their LBCC values and other prediction measures; second, we selected the top 100, 200, 300, 400, 500, and 600 proteins as essential proteins; and finally, the number of true essential proteins was determined. For the YMIPS dataset, LIDC, the most recent method, had the best performance, with 66, 124, 177, 224, 265, and 314 true essential proteins identified at six levels from the top 100 to top 600. By comparison, the numbers of true essential proteins predicted by LBCC were 75, 145, 199, 248, 305, and 343, respectively. Compared to LIDC, LBCC exhibited superior performance and increased the prediction precision by more than 13, 16, 12, 10, 15 and 9 percent at six levels from the top 100 to top 600.

For the YMBD dataset, except for LBCC, the largest numbers of true essential proteins identified were 43 (BC), 90 (BC), 133 (BC), 174 (BC), 212 (BC), and 258 (NC) at six levels from the top 100 to top 600. By comparison, LBCC identified 65, 120, 176, 231, 275, and 315 true essential proteins, improving the prediction precision by more than 51, 33, 32, 32, 29, and 22 percent at six levels from the top 100 to top 600.

For the YHQ dataset, BC achieved the best result at the top 100 level, and SC and EC attained the best results at the top 200 level. At four levels from the top 300 to top 600, LBCC produced the best results, and the numbers of true essential proteins identified were 169, 241, 296 and 348. For the YDIP dataset, LIDC achieved the best results at the top 100, 200, 300 and 500 levels, and LBCC attained the best results at the top 400 and 600 levels. At six levels, the numbers of true essential proteins identified by LIDC were 76, 152, 209, 260, 313, and 354. By comparison, the numbers of true essential proteins identified by LBCC were 74, 135, 205, 262, 308, and 361, respectively. The results predicted by LBCC were similar to those obtained using LIDC at the top 100, 300 and 500 levels.

Thus, our experiments indicate that LBCC can identify more essential proteins than the other methods in most cases.

Results on human PPI network

To further evaluate the performance of the proposed method LBCC, we also applied it to identify essential proteins on a human PPI network. The human PPI network data marked HDIP were from the DIP database, the essential proteins were collected from DEG, and the protein complex set marked HCOM was from CORUM (Comprehensive Resource of Mammalian protein complexes). HDIP consisted of 4647 interactions and 2914 proteins, including 1887 essential proteins, and HCOM contained 1283 protein complexes. We compared the performances of LBCC and the other seven methods in six levels from the top 100 to top 600. Almost every method achieved more than 70 percent precision due to the large proportion of essential proteins, and LBCC achieved the best results at the top 100-400 levels. However, LBCC tended to provide less desirable results compared with LIDC at the top 500 and 600 levels.

CONCLUSION

The identification of essential proteins is helpful for comprehending the minimal requirements for cellular life, and many approaches based on topological properties have been proposed for discovering essential proteins in PPI networks. Most of the topology-based methods only concentrate on either local or global characteristics and are also sensitive to the network structure. LIDC outperformed classical topological centrality measures. In this paper, we propose a new method, LBCC, based on the combination of three characteristics of the protein-protein interaction network, i.e., Den1(v), Den2(v), BC(v) and IDC(v), which represent both local and global characteristics and information on protein complexes. We applied LBCC to four PPI networks of Saccharomyces cerevisiae: YMIPS, YMBD, YHQ and YDIP. We then conducted comprehensive comparisons of LBCC and the other seven previously proposed methods, including DC, BC, SC, EC, NC, LAC and LIDC, in terms of the number of true essential proteins identified. At the six levels from the top 100 to top 600, LBCC outperformed recent prediction methods on the YMIPS and YMBD datasets. In particular, LBCC improved the prediction precision by more than 10 percent compared to LIDC. Moreover, we also applied LBCC to a human PPI network, HDIP. The experimental results show that LBCC is also effective for predicting essential proteins for the HDIP network. Hence, we conclude that LBCC is a more effective method for predicting essential proteins, occasionally significantly. In future studies, we will integrate additional information, such as domain information, gene ontology and gene expression data, to predict essential proteins more effectively and accurately.

REFERENCE

1. Qin C, Sun Y, Dong Y (2016) A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes. PLoS ONE 11(8): e0161042. doi:10.1371/journal. Pone.0161042
2. https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/protein-protein-interaction-networks
3. https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction
4. https://www.nature.com/subjects/protein-protein-interaction-networks
5. Luo J, Qi Y (2015) Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes. PLoS ONE 10(6): e0131418. doi:10.1371/journal. pone.0131418