A Dynamic Weighted PPI Network-Based Essential Protein Identification Algorithm Postprint

Authors: Yang Shuxin, Lu Jihua, Tang Darong

Date: 2018-05-20T00:00:00+00:00

Abstract

Compared with static PPI networks, dynamic PPI networks better reflect the actual situation of protein-protein interactions and effectively reduce false negatives in PPI networks. Existing essential protein prediction methods are primarily applied to static PPI networks, neglecting the dynamic characteristics of PPI networks. To effectively predict essential proteins, we utilize gene expression data to extract dynamic information of proteins, combine it with static PPI networks to construct dynamic PPI networks, then introduce GO terms to weight the network, and propose a new prediction method—DWE—based on the dynamic weighted PPI network. This method measures the essentiality of a protein in the network by the ratio of the sum of its dynamic weighted edges to the number of times the protein appears in the dynamic network. Experimental results demonstrate that dynamic weighted PPI networks help improve the prediction accuracy of essential proteins, and the DWE method outperforms several other essential protein prediction methods.

Full Text

A Novel Algorithm for Essential Protein Identification Based on Dynamic Weighted PPI Networks

Yang Shuxin, Lu Jihua, Tang Darong
School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China

Abstract

Compared with static PPI networks, dynamic PPI networks better reflect the real interactions among proteins and effectively reduce false negatives in PPI data. Most existing essential protein prediction methods are applied to static

PPI networks, neglecting the inherent dynamic characteristics of protein interactions. To effectively predict essential proteins, we extract dynamic information from gene expression data and integrate it with static PPI networks to construct dynamic PPI networks. We then introduce GO term similarity for network weighting and propose a new prediction method called DWE (Dynamic Weighted Essentiality). This method measures the importance of a protein by calculating the ratio between the sum of its dynamic weighted edges and the number of temporal networks in which the protein appears. Experimental results demonstrate that dynamic weighted PPI networks improve prediction accuracy, and the DWE method outperforms several other essential protein prediction approaches.

Keywords: dynamic network; essential proteins; GO terms; dynamic weighted PPI network

0 Introduction

Essential proteins are indispensable for life activities. Identifying essential proteins not only aids in understanding biological processes but also holds significant importance for drug design and disease treatment. While biological researchers have traditionally identified essential proteins through experiments, these approaches are costly, inefficient, and applicable only to certain organisms. Consequently, computational methods have been developed to improve prediction efficiency and reduce costs.

Existing essential protein prediction methods can be categorized based on the type of PPI network used: static PPI network-based approaches and dynamic PPI network-based approaches. Static PPI network methods primarily include topological feature-based predictors such as Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Subgraph Centrality (SC), Eigenvector Centrality (EC), Information Centrality (IC), Local Average Connectivity (LAC), and Neighbor Centrality (NC). Additionally, several methods integrate biological information, including PeC, DWC, and UDoNC. In 2016, Cui et al. introduced the LDA model into the existing CPPK framework for essential protein prediction, while Yang et al. combined edge clustering coefficients with random walk models for identification purposes.

Currently, research on dynamic PPI networks mainly focuses on mining protein complexes and functional modules, with relatively few studies addressing essential protein detection methods. In 2015, Xiao et al. constructed dynamic PPI networks based on time-series models and applied six existing essential protein prediction methods to these networks. Although this approach considered the impact of network dynamics on prediction results, it neglected the influence of false positives and false negatives in PPI data. To mitigate these effects, this paper constructs a novel dynamic weighted PPI network (DWPPIN) by incorporating both dynamic characteristics and functional similarity, and proposes a new detection method called DWE based on this network.

1 Essential Protein Prediction Based on Dynamic Weighted PPI Networks

The DWE method follows a three-stage workflow: First, we use static PPI networks as a framework and construct dynamic PPI networks from gene expression profiles. Second, for each temporal network within the dynamic PPI network, we assess interaction reliability using both edge clustering coefficients and functional similarity between proteins. Finally, we calculate protein scores in the dynamic weighted network to determine their importance.

1.1 Dynamic PPI Network Construction

Dynamic PPI networks can be constructed based on protein co-expression patterns. Co-expression occurs when two interacting proteins are both active simultaneously. A protein is considered active when its expression level exceeds a certain threshold. Since different proteins exhibit varying expression levels, a universal threshold is inappropriate. Following established practice, we use each protein' s average expression level as its individual threshold. When a protein' s expression value at a given time point exceeds its average, we consider it active.

The gene expression profile contains expression level data across three metabolic cycles (12 time points per cycle, totaling 36 time points). The average expression level for each protein is calculated as:

$$\text{avg}_v = \frac{1}{36}\sum_{t=1}^{36} expr(v,t)$$

where $expr(v,t)$ represents the expression value of protein v at time t, and avg_v represents its average expression value.

After calculating average expression levels, we determine each protein' s active status at each time point based on the gene expression profile. For an interaction (u,v) in the static PPI network, if both proteins u and v are active at time t, then the interaction exists at that time. This can be expressed as:

$$e(u,v,t) = \begin{cases} 1 & \text{if } expr(u,t) > \text{avg}_u \text{ and } expr(v,t) > \text{avg}_v \\ 0 & \text{otherwise} \end{cases}$$

In Equation (2), $e(u,v,t) = 1$ indicates that interaction (u,v) exists at time t; otherwise, it does not exist.

This approach yields 36 distinct temporal PPI networks, each containing different active proteins and interactions. The dynamic PPI network can be formally defined as:

Definition 1 Dynamic PPI network $DG = \{G_1, G_2, ..., G_l\}$, where l represents the number of time points. $G_i = \{V_i, E_i\}$ is the PPI network at time i, where

$V_i = \{v_{i1}, v_{i2}, ..., v_{in_i}\}$ represents the set of active proteins at time i, and $E_i = \{e_{i1}, e_{i2}, ..., e_{im_i}\}$ represents the set of protein interactions at time i.

Figure 1 [Figure 1: see original paper] illustrates the dynamic network construction process. In the static PPI network, proteins A and B interact and are both active at time points 1, 3, and 4. Therefore, the A-B interaction appears only at these three time points. Although proteins E and C interact in the static network, they are never active simultaneously, so their interaction never appears in the dynamic network. Conversely, while proteins F and D are both active at time 7, they do not interact in the static network, so no interaction exists at that time point.

1.2 Dynamic PPI Network Weighting

Analysis of essential proteins reveals that they tend to cluster together, and their essentiality is closely related to interaction reliability. We simultaneously consider both topological and biological features, using GO semantic similarity and edge clustering coefficients to measure interaction reliability.

Interaction reliability can be evaluated through functional similarity between proteins. Higher functional similarity indicates more reliable interactions. GO terms provide structured natural language annotations for gene functions. Research has demonstrated that proteins sharing more GO terms exhibit greater functional similarity. Therefore, interaction reliability can be assessed through functional similarity based on GO terms.

Let $GO(v)$ denote the set of GO annotations for protein v, and $GO\_sim(u, v)$ represent the GO semantic similarity between proteins u and v. The functional similarity is calculated as:

$$GO\_sim(u, v) = \frac{|GO(u) \cap GO(v)|}{|GO(u) \cup GO(v)|}$$

where $|GO(u) \cap GO(v)|$ represents the number of shared GO terms, and $|GO(u) \cup GO(v)|$ represents the total number of distinct GO terms.

Essential proteins are not only associated with interaction reliability but also tend to cluster together. The edge clustering coefficient serves as an important topological feature of PPI networks, describing interaction importance and assessing the probability that proteins belong to the same cluster, while effectively identifying essential proteins. The edge clustering coefficient is calculated as:

$$ECC(u, v) = \frac{\text{num}_{\triangle}(u, v)}{\min(d(u) - 1, d(v) - 1)}$$

where $\text{num}_{\triangle}(u, v)$ represents the number of triangles formed by proteins u and v, and $d(u)$ and $d(v)$ represent the degrees of proteins u and v, respectively.

Definition 2 Dynamic weighted PPI network $DWG = \{G_1, G_2, ..., G_l\}$, where l represents the number of time points, and $G_i = \{V_i, E_i, WE_i\}$ is the weighted PPI network at time i. Here, $V_i = \{v_{i1}, v_{i2}, ..., v_{in_i}\}$ represents the set of active proteins at time i, $E_i = \{e_{i1}, e_{i2}, ..., e_{im_i}\}$ represents the set of interactions at time i, and $WE_i = \{we_{i1}, we_{i2}, ..., we_{im_i}\}$ represents the set of interaction weights.

The weight of interaction (u, v) in temporal network G_i is calculated as:

$$WE_i(u, v) = GO\_sim(u, v) + ECC(u, v)$$

1.3 Essential Protein Identification

The essentiality of a protein is determined by its score in the dynamic weighted PPI network constructed in Section 1.2. Since active proteins differ across temporal networks, we consider the number of temporal networks containing each protein when calculating scores. The DWE score for a protein is defined as the ratio between the sum of its weighted edges across all temporal networks and the number of networks in which it appears:

$$DWE(v) = \frac{\sum_{i=1}^{l} \sum_{u \in N_i(v)} WE_i(v, u)}{freq(v)}$$

where $N_i(v)$ represents the neighbor set of protein v in temporal network G_i, $WE_i(v, u)$ represents the weight of edge (v, u) in G_i, and $freq(v)$ represents the number of temporal networks containing protein v.

We rank all proteins by their DWE values in descending order. Proteins with higher DWE scores are more likely to be essential. The top K proteins are output as candidate essential proteins predicted by DWE, where K is the number of essential proteins in the dynamic weighted PPI network.

The DWE algorithm is described as follows:

DWE Algorithm
Input: Static PPI network G, gene expression data GED, GO annotation GOA, K
Output: Top K ranked proteins by DWE value in descending order

1. Construct dynamic PPI network based on GED according to formulas (1) and (2)
2. For each $G_i \in DG$:
    - For each $(u, v) \in E_i$:
        - Compute $GO\_sim(u, v)$
        - Compute $ECC(u, v)$
        - Compute $WE_i(u, v) = GO\_sim(u, v) + ECC(u, v)$

- End for
- For each $v \in V_i$:
  - Compute the sum of weighted edges connected to v
- End for

3. End for
4. a) Compute the DWE score for all nodes using formula (6)
5. b) Return the top K ranked proteins according to their DWE values

2 Experimental Results and Analysis

2.1 Experimental Data

Due to their high completeness and reliability, yeast protein interaction networks and essential protein datasets are used to validate our method. The experimental datasets are described below:

1. **Yeast Protein-Protein Interaction Network:** Obtained from the DIP database, containing 5,093 proteins and 24,743 interactions.
2. **Reference Essential Proteins:** Integrated from four databases (MIPS, SGD, DEG, and SGDP), containing 1,285 essential proteins, of which only 1,167 appear in the yeast PPI network.
3. **Yeast Gene Expression Profiles:** Downloaded from NCBI Gene Expression Omnibus (dataset GSE3431), stored as an $m \times n$ matrix where each entry represents a protein' s expression level at a specific time point. The dataset contains expression data for 6,740 gene products, with 4,981 present in the yeast PPI network.
4. **Yeast Protein GO Annotations:** Downloaded from the Gene Ontology database (version dated December 24, 2016), including three categories: biological process, cellular component, and molecular function.

2.2 Experimental Results

2.2.1 Impact of PPI Network Type on Detection Performance Since DC, LAC, and NC measure protein importance from different perspectives (neighbor count, neighbor importance, and edge importance, respectively), we selected these three methods for evaluation. Figure 2 [Figure 2: see original paper] shows the number of essential proteins identified by these methods across three different PPI networks. The results demonstrate that dynamic weighted networks significantly improve detection performance compared to static PPI networks. When selecting the top 600 candidate proteins, the accuracy of dynamic weighted PPI networks increases by 20%. Although the identification accuracy of dynamic weighted and dynamic PPI networks is similar for small candidate sets, the dynamic weighted network gradually outperforms the dynamic network as the candidate set expands. These results confirm that dynamic weighted PPI networks enhance essential protein prediction.

2.2.2 Rank-and-Filter Analysis The rank-and-filter approach calculates a specific metric for each protein, ranks proteins in descending order, selects top-ranked proteins as candidates, and compares them against known essential proteins. In this experiment, we selected proteins ranked in the top 1%, 5%, 10%, 15%, 20%, and 25% as candidate essential proteins.

Figure 3 [Figure 3: see original paper] presents the identification results for 11 methods. Among the seven classical centrality methods, NC achieves the best performance. Compared to NC, DWE improves identification accuracy by 28.12%, 15.72%, 15.90%, 11.52%, 8.41%, and 6.62% at different candidate percentages. When compared to biology-integrated methods (PeC, WDC, UDoNC), DWE maintains a clear advantage when candidate essential proteins constitute no more than 25% of the total. Although DWE' s advantage diminishes as the candidate set expands, it still shows improvement at the 25% threshold, with accuracy gains of 8.21%, 2.47%, and 1.22% over PeC, WDC, and UDoNC, respectively.

For more detailed performance evaluation, we employed the jackknife method to compare DWE against ten other centrality measures. In Figure 4 [Figure 4: see original paper], the X-axis represents the cumulative number of candidate essential proteins, and the Y-axis represents the cumulative number of true essential proteins. The area under the jackknife curve indicates each method' s performance, with larger areas corresponding to higher identification rates. Figure 4 shows that DWE achieves the best identification rate compared to DC, BC, EC, SC, IC, NC, and CC. As shown in Figure 4(a), the DWE curve lies clearly above the PeC curve, indicating superior performance. Figure 4(b) demonstrates that DWE also outperforms WDC and UDoNC.

Additionally, we evaluated each method using six metrics: Sensitivity (SN), Specificity (SP), F-measure, Accuracy (ACC), Positive Predictive Value (PPV), and Negative Predictive Value (NPV), calculated as follows:

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

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

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

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

$$F\text{-measure} = \frac{2 \times SN \times PPV}{SN + PPV}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Based on the number of known essential proteins, we selected the top 1,167 ranked proteins as predictions and calculated the six metrics for each method. Table 1 compares DWE with ten other methods across all six metrics, showing that DWE achieves the highest values for every metric.

3 Conclusion

This paper leverages the dynamic characteristics of PPI networks and integrates biological features to construct a dynamic weighted PPI network, proposing the DWE method for essential protein identification. By reducing data noise and improving network reliability, DWE significantly enhances prediction accuracy compared to existing methods. The experimental results demonstrate that DWE outperforms both classical centrality measures and biology-integrated approaches across multiple evaluation metrics.

References

[1] Jeong H, Oltvai Z N, Barabási A-L. Prediction of protein essentiality based on genomic data [J]. ComPlexUs, 2002, 1 (1): 19-28.

[2] Jeong H, Mason S P, Barabási A-L, et al. Lethality and centrality in protein networks [J]. Nature, 2001, 411 (6833): 41-42.

[3] Joy M P, Brock A, Ingber D E, et al. High-betweenness proteins in the yeast protein interaction network [J]. BioMed Research International, 2005, 2005 (2): 96-103.

[4] Wuchty S, Stadler P F. Centers of complex networks [J]. Journal of Theoretical Biology, 2003, 223 (1): 45-53.

[5] Estrada E, Rodriguez-Velazquez J A. Subgraph centrality in complex networks [J]. Physical Review E, 2005, 71 (5): 056103.

[6] Bonacich P. Power and centrality: A family of measures [J]. American journal of sociology, 1987: 1170-1182.

[7] Stephenson K, Zelen M. Rethinking centrality: Methods and examples [J]. Social Networks, 1989, 11 (1): 1-37.

[8] Li Min, Wang Jianxin, Chen Xiang, et al. A local average connectivity-based method for identifying essential proteins from the network level [J]. Computational Biology & Chemistry, 2011, 35 (3): 143-50.

[9] Wang Jian, Li Min, Wang Huan, et al. Identification of Essential Proteins Based on Edge Clustering Coefficient [J]. IEEE//ACM Transactions on Computational Biology & Bioinformatics, 2012, 9 (4): 1070.

[10] Li Min, Zhang Hanhui, Wang Jianxin, et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data [J]. BMC Systems Biology, 2012, 6 (1): : 15.

[11] Tang Xiwei, Wang Jianxin, Zhong Jiancheng, et al. Predicting Essential Proteins Based on Weighted Degree Centrality [J]. IEEE//ACM Trans Comput Biol Bioinform, 2014, 11 (2): 407-18.

[12] Peng Wei, Wang Jianxin, Cheng Yingjiao, et al. UDoNC: An Algorithm for Identifying Essential Proteins Based on Protein Domains and Protein-Protein Interaction Networks [J]. IEEE/ACM Trans Comput Biol Bioinform, 2015, 12 (2): 276-88.

[13] Xiao Qianghua, Wang Jianxin, Peng Xiaoqing, et al. Identifying essential proteins from active PPI networks constructed with dynamic gene expression [J]. BMC Genomics, 2015, 16 (3): 1.

[14] Cui Xin, Shao Mingyu. Identification of essential proteins combining topic features and topological properties of interaction networks [J]. Computer Applications and Software, 2016, 33 (08): 283-288.

[15] Yang Liping, Lu Songfeng, Huang Yu. A key protein prediction method based on random walk model [J]. Journal of Huazhong Agricultural University, 2016, 35 (6): 86-91.

[16] Wang Jianxin, Peng Xiaoqing, Li Min, et al. Construction and application of dynamic protein interaction network based on time course gene expression data [J]. Proteomics, 2013, 13 (2): 301-312.

[17] Wang Huan, Li Min, Wang Jianxin, et al. A New Method for Identifying Essential Proteins Based on Edge Clustering Coefficient [J]. Lecture Notes in Computer Science, 2011, 6674: 87-98.

[18] Xenarios I, Salwínski Ł, Duan X J, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions [J]. Nucleic Acids Research, 2002, 30 (1): 303.

[19] Hw M, D F, Kf M, et al. MIPS: analysis and annotation of proteins from whole genomes [J]. Nucleic Acids Research, 2006, 34 (2): 169-72.

[20] Cherry J M, Adler C, Ball C, et al. SGD: Saccharomyces Genome Database [J]. Nucleic Acids Research, 1998, 26 (1): 73-9.

[21] Zhang Ren, Lin Yan. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes [J]. Nucleic Acids Research, 2009, 37 (Database issue): D455-D458.

[22] Saccharomyces Genome Deletion Project [EB/OL]. http://www-sequence.stanford.edu/group/yeast_deletion_project.

[23] Tu B P, Kudlicki A, Rowicka M, et al. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes [J]. Science, 2005, 310

(5751): 1152-1158.

[24] Consortium T G O. Gene Ontology Consortium: going forward [J]. Nucleic Acids Research, 2015, 43 (Database issue): 1049-56.

Note: Figure translations are in progress. See original paper for figures.