OXFORD

Briefings in Bioinformatics, 21(2), 2020, 486–497

doi: 10.1093/bib/bbz011
Advance Access Publication Date: 12 February 2019
Review article

Computational methods for identifying the critical nodes in biological networks

Xiangrong Liu, Zengyan Hong, Juan Liu, Yuan Lin, Alfonso Rodríguez-Patón, Quan Zou and Xiangxiang Zeng

Corresponding author: Quan Zou, Insitute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610000, Chengdu, China. E-mail: zouquan@nclab.net; Xiangxiang Zeng, Department of Computer Science, Xiamen University, 361005, Xiamen, China. E-mail: xzeng@xmu.edu.cn

Abstract

A biological network is complex. A group of critical nodes determines the quality and state of such a network. Increasing studies have shown that diseases and biological networks are closely and mutually related and that certain diseases are often caused by errors occurring in certain nodes in biological networks. Thus, studying biological networks and identifying critical nodes can help determine the key targets in treating diseases. The problem is how to find the critical nodes in a network efficiently and with low cost. Existing experimental methods in identifying critical nodes generally require much time, manpower and money. Accordingly, many scientists are attempting to solve this problem by researching efficient and low-cost computing methods. To facilitate calculations, biological networks are often modeled as several common networks. In this review, we classify biological networks according to the network types used by several kinds of common computational methods and introduce the computational methods used by each type of network.

Key words: essential proteins; essential genes; computational methods; biological networks; topological features; biological information

Introduction

Bioinformatics is one of the core areas of natural science in the 21st century. Some studies in this field include [1–23]. Biological networks have also become a biological research hotspot because they provide a graph of the whole cell and the entire organism, thereby enabling researchers to systematically study a large number of collective biological behaviors and co-expressed features. Nodes in biological networks represent proteins, genes, RNA and DNA, among others. Edges in networks correspond to physical, biochemical or functional interactions among nodes. Different biological networks can be established based on different interactions [24]. Common biological networks include protein–protein interaction (PPI), metabolic, gene regulatory and signal transduction networks. Individual proteins constitute, through their interaction, a PPI network to participate in various aspects of life processes, such as biological signal transmission, gene expression regulation, energy and substance metabolism and cell-cycle regulation. Each node in a PPI network represents a protein, and an edge between two points represents the relationship between them. A metabolic network refers to a network of metabolic reactions and regulatory mechanisms

Xiangrong Liu is a professor at Xiamen University. His research interests include bioinformatics and data mining.
Zengyan Hong is a graduate student at Xiamen University. Her research interest is classification of proteins in bioinformatics.
Juan Liu is an associate professor at Xiamen University. Her research interests include industrial automation and embedded systems.
Yuan Lin is a senior engineer at DNB Bank ASA.
Alfonso Rodríguez-Patón is an associate professor at Xiamen University. His research interests include bio-computing and bioinformatics.
Quan Zou is a professor at the University of Electronic Science and Technology of China. His research interest is bioinformatics.
Xiangxiang Zeng is an associate professor at Xiamen University. His research interests include bio-computing and bioinformatics.
Submitted: 25 October 2018; **Received (in revised form):** 3 December 2018

that control these responses, which describe intracellular metabolism and physiological processes. A gene regulatory network is formed by the interactions between genes and genes in cells (or within a particular genome). A signal transmission network comprises molecules and enzymes involved in the signal transduction pathway and the biochemical reactions that occur between them [25].

Essential genes and essential proteins are related to the survival of organisms. The identification of essential genes and proteins helps elucidate the minimum requirements of cell survival and has practical purposes such as drug design [26]. Thus, the search for essential genes and proteins has attracted increased attention. Previous researchers have often used experimental methods of identifying essential genes and proteins, such as single-gene knockouts, RNA interference and conditional knockouts. However, these methods require much time, manpower and money [27], so computational methods of identifying essential genes and proteins is important. Indeed, efficient and low-cost computational methods can quickly and accurately identify essential genes and proteins in biological networks.

Moreover, the construction of networks with specific characteristics for the nature of different biological networks can help improve the accuracy and efficiency of computational methods. For example, a sequential relationship exists between transcription factors and regulated genes in regulatory networks; thus, regulatory networks are generally modeled as directed networks [28]. To increase the reliability of networks, some specific methods model these networks as dynamic and weighted by combining biological information.

In this review, we classify biological networks according to the network types used by several common computational methods and introduce the computational methods used by each type of network. The rest of this review is summarized as follows. In Sections In undirected network, In directed network, In weighted network and In dynamic network, we introduce the computational methods used to find key nodes in undirected, directed, weighted and dynamic networks, respectively. In Section Summary of the methods mentioned, we summarize the methods mentioned in this review. In Section Biological network data sets, we describe the data acquisition method for various biological networks and some instances of biological networks. Finally, we conclude this review with a summary.

In undirected network

We define an undirected graph as $G = (V, E)$, whose V is a finite set of nodes that represents the biomolecules in the network, and E is a finite set of edges between the nodes. The edge connecting two nodes is undirected. A symmetric adjacency matrix A describes an undirected graph, G. A is a $(|V| \times |V|)$ matrix, where $a_{ij} = 1$ if and only if $(i, j) \in E$; otherwise, $a_{ij} = 0$. Many biological networks, such as the PPI network, which do not consider the direction of the action, are often built as undirected networks. In this section, we introduce the computational methods used in undirected networks.

Methods based on topological features

Typical centrality measures are usually used in complex networks to identify the key nodes. Degree centrality (DC) is the simplest indicator in characterizing the importance of nodes in a network. DC thinks that the greater the number of neighbors of a node, the greater its influence will be. Eccentricity centrality (EC) is also often used to rank the importance of nodes in a network. The eccentricity of node v in network G is the maximum value of the distance from v to other nodes. The more the EC of the node tends to the radius of the network, the closer the node is to the center of the network. The more the EC of the node tends to the diameter of the network, the closer the node is to the edge of the network. Closeness centrality (CC) calculates the importance of a node by means of the average of the shortest path from one node to another. Eigenvector Centrality thinks that the importance of a node depends not only on the number of its neighbors (the degree of the node) but also on the importance of each neighbor. Koschützki *et al.* [29] discussed these central measures and showed their suitability in PPI networks and transcriptional regulation networks.

Betweenness centrality (BC), defined as the number of shortest paths through the current node in a network, is also a kind of typical centrality measure. BC portrays the control of a node on the network flow transmitted along the shortest path in the network. Based on the analysis of the betweenness measure, Joy *et al.* [30] found that proteins with a higher betweenness are more likely to be critical, and the evolutionary age of the protein is positively correlated with the value of betweenness.

Subgraph centrality (SC), denoting the number of nodes appearing in different subgraphs, considers the importance of the contribution of the nodes in the global loop; the shorter the loop, the greater the contribution. Estrada [31] used SC to identify the essential protein in yeast PIN and found that SC performs better than DC, BC, CC and EC.

Wang *et al.* [32] presented a new method, SoECC, based on edge clustering coefficients to identify essential proteins. This method effectively binds the features of edges and nodes. The experimental results of the yeast PPI network show that the number of essential proteins discovered by SoECC is generally higher than DC, BC, CC, SC, EC, etc.

Many centrality methods focus only on the location of proteins, ignoring the relationship between protein interactions (edge features). Wang *et al.* [33] proposed a new centrality method, neighborhood centrality (NC), based on edge clustering coefficients. NC not only considers the centrality of nodes but also considers the relationship between nodes and neighboring nodes. The authors used the NC method for the yeast PPI network and found that NC performed better than the other methods, such as DC, BC, CC, etc. Also, the NC method is suitable for different types of networks.

Many essential proteins are not highly connected. By analyzing these proteins, Li *et al.* [34] found that their neighbors had little interaction. Therefore, they developed a new local method, local average connectivity (LAC), to determine the essentiality of proteins by evaluating the relationship between proteins and their neighbors. They used the yeast PPI networks, obtained from two different databases, to validate LAC. The experimental results of these two networks show that LAC performs better than DC, BC, CC, EC, etc.

In 2017, Li *et al.* [35] proposed a new method, ME, which studied the topological characteristics of key proteins from the perspective of network motif. Unlike previous centrality metrics, this approach considers both the centrality of the node and the relationship between the node containing the topic and its neighbors. The necessity of a node is determined by the sum of the density of the nearest neighbor extensions of all the motifs that contain it. The experimental results show that this method is superior to traditional topology measures: DC, BC, CC, SC, EC, information centrality (IC), LAC, NC and MC.

Methods based on combination of topological features and biological information

As mentioned above, there are many topology-based computational methods. However, these methods, which contain many false positives and false negatives, are very sensitive to the robustness of the network. Therefore, more and more researchers are beginning to study methods in integrating topology information and biological information.

Considering that many species already have many known essential proteins, Li *et al.* [36] proposed an a priori knowledge-based scheme to discover new essential proteins from PPI networks. The authors found that most of the essential proteins are neighbors of other essential proteins that tend to be in the same cluster and are co-expressed. Based on the new scheme, the authors developed two essential protein discovery algorithms: CPPK and CEPPK. CPPK predicts new essential proteins based on network topology. CEPPK detects new essential proteins by integrating network topology and gene expression. The PPI network, based on *Saccharomyces cerevisiae*, verified the performance of CPPK and CEPPK. The experimental results show that a priori knowledge of the known essential protein is effective in improving the prediction precision.

SoECC, a method that has good performance, is based on topology alone. However, SoECC cannot identify essential proteins that are rarely linked to other proteins. Therefore, Ma *et al.* [37] put forward a new centrality algorithm, CSC, in identifying essential proteins. CSC integrates the topological features of the PPI network and the in-degree of proteins in the complex. They used the CSC algorithm to identify proteins in the *S. cerevisiae* PPI network. The results show that the proportion of essential proteins identified by the CSC algorithm is higher than the results using the other 10 centrality methods: DC, BC, CC, SC, EC, IC, Bottle Neck (BN), LAC, SoECC and PeC.

Li *et al.* [38] proposed a new method, GOS, to identify essential proteins by integrating gene expressions, orthology and subcellular localization information. The gene expressions and subcellular localization information are used to determine whether the neighbors in the PPI network are reliable. When analyzing the topological features of proteins in a PPI network, only reliable neighbors are considered. The experimental results on the yeast PPI network show that the proposed method, GOS, is superior to the existing methods: DC, BC, CC, SC, EC, NC, CSC, etc.

Topological centrality measures do not perform well in single protein data sets, and multi-information fusion measures are sensitive to the topology of the network. Therefore, Qi *et al.* [39] created a novel topology centrality measure, local interaction density (LID), for predicting essential proteins based on local interaction density. Different from the previous method, LID obtains the importance of nodes from the interaction density among its neighbors by topological analysis of the real protein in the protein complexes set from the perspective of the biological module. Thus, LID has certain biological rationality and is insensitive to the topology of different networks.

Zhang *et al.* [40] proposed a new method for predicting key proteins by integrating Gene expression profiles and Gene Ontology (GO) annotation data, named GEG. This method solves the problem of unsatisfied precision when predicting a small number of essential proteins, and the method has good robustness in the case of disturbance. The authors evaluated GEG on two PPI networks of *S. cerevisiae*, which achieved better performance than DC, BC, NC, etc.

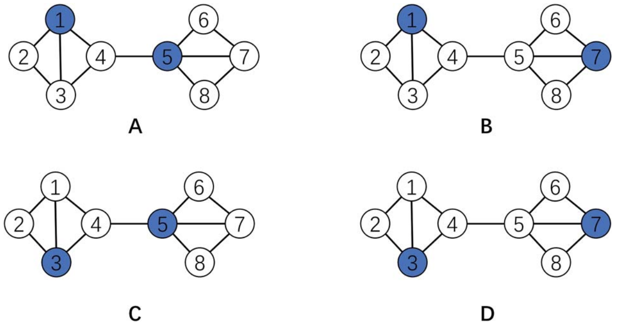

Figure 1. An example of detecting the driver nodes in an undirected network. The concept of an MDS is illustrated in this figure. An MDS is an optimized sub-set of biomolecules (blue nodes) from which each remaining biomolecule (white node) can be reached by one step. Blue nodes are driver nodes. We can see different optimization methods would generate different MDS configurations (there are four MDSs in the network: A: {1,5}; B: {1,7}; C: {3,5}; D: {3,7}), and it is difficult to evaluate which one is better.

Methods based on system controllability

Due to the interaction between biomolecules in the biomolecular network, disrupting some biomolecules may affect other biomolecules, which may cause the entire network to change state and eventually change cell behavior. Therefore, controlling cell behavior by regulating certain biomolecules in the network is one of the most concerned issues in systems biology. The research on network controllability is to find the fewest nodes in the network, and exerting control on these nodes can make the network run to our desired target state. A type of these nodes, Driver Nodes, which are controlled by different signals, can offer full control over the network. From the perspective of network controllability, indispensable proteins and genes can be seen as the driver nodes. Some researchers have developed many computational methods to find the driver nodes.

Many models are used to find driver nodes. However, the existing minimum dominating set (MDS) models have a large problem in that they generate different MDSs when using different optimization methods (Figure 1). Therefore, the true set of protein driver nodes cannot be found. To solve this problem, Zhang *et al.* [41] created a centrality-corrected minimum dominating set (CC-MDS) model, which includes the heterogeneity in DC and BC of proteins, based on the theory that a node, which has high degrees and high betweenness is more likely to be the controller. Experimental results show that CC-MDS performs better than the previous standard MDS model.

Methods based on machine learning

With the rapid growth of biological information, biological science technology has greatly enriched the biological science data resources. Machine learning, having the ability to learn from data and experiences, can extract knowledge from a large amount of biological data. Finding essential genes and proteins in biological networks can be considered as a problem of node classification. Figure 2 shows the workflow of machine learning-based methods for critical genes and proteins prediction. The advantages of machine learning–based computing methods are efficiency and accuracy.

Plaimas *et al.* [42] proposed an integrated machine learning approach that combines topological-based methods with genetic information to predict essential genes when no experimental knock-out data is available. This method uses detailed features of network topology, sequence information

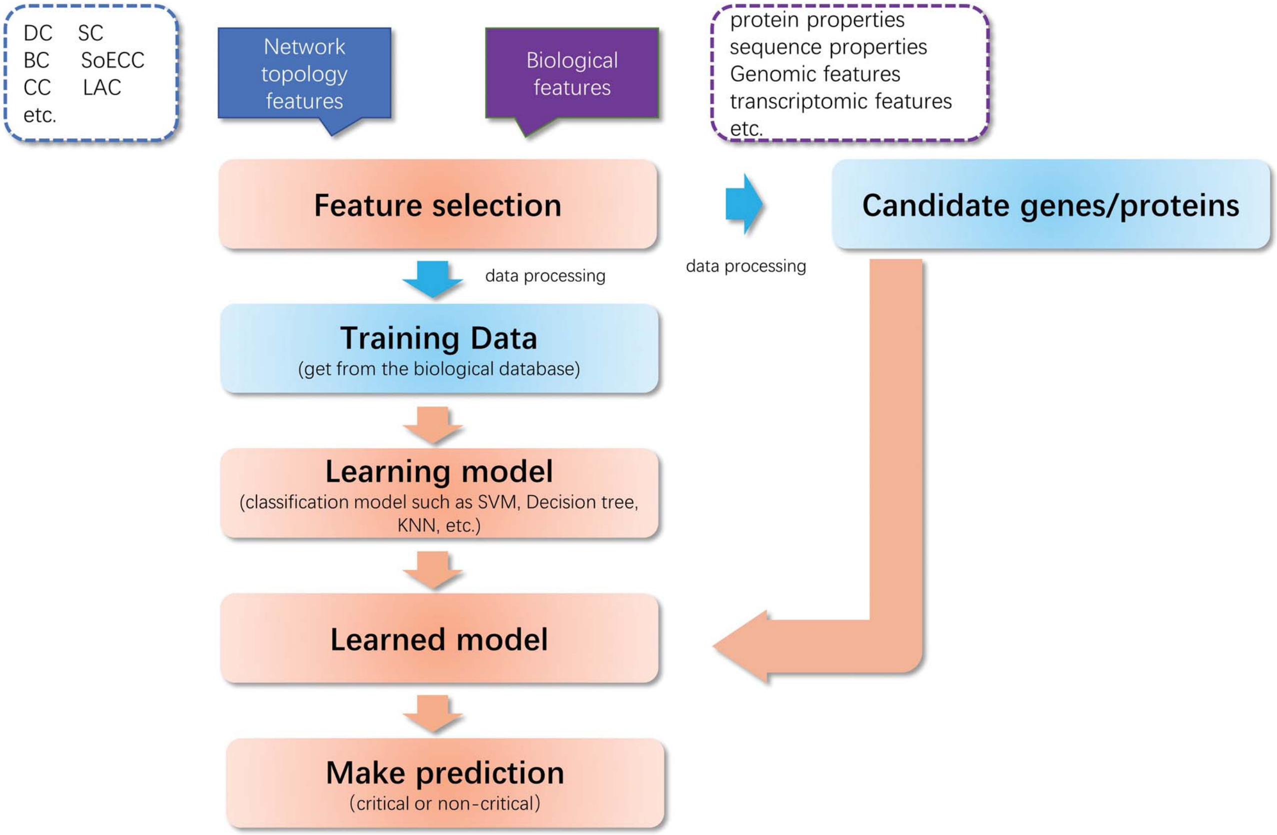

Figure 2. The workflow of machine learning–based methods for critical genes and proteins prediction.

and microarray data to construct a classifier: support vector machines (SVM). Then the information of the essential genes obtained through the experimental whole-genome screen was used as training sets to train the classifier. Finally, the trained classifier is used in the query organism to predict the criticality of each gene. This method has found eight targets more than the experimental method. This prediction method does not depend on the information of the essential genes of the query organism but solely on the characteristics that were calculated from the metabolic network and genomic and transcriptomic information of the query organism. When the experimental data on a whole gene knockout screen does not apply to the organism to be predicted, this method is useful for inferring drug targets.

In previous studies, more or fewer of the topology and protein properties were used without considering some basic properties, such as sequence and protein chemistry. Therefore, Hor *et al.* [43] used the improved reverse feature selection method to select topological properties, protein properties, sequences and other basic properties as feature sets. Then, using the selected features, they constructed an SVM model with various feature subsets, which is improved by LIBSVM and selects radial basis function as the core function. They used cross-validation to verify the effect of the improvements. The final results show that the N9 model (nine features) performs better than subsequent models.

In order to solve the problem of the unbalanced training set, difficult parameter selection and algorithm limitation in machine learning, Nandi *et al.* [44] proposed a simple but powerful SVM-based learning strategy. The authors selected gene and protein sequences, gene expression and network topological and flux-based features for *Escherichia coli* K-12 MG1655 metabolism as features. With the chosen training characteristics, the method is capable of capturing the minimal set of essential genes that are proven necessary in any given environment.

In directed network

Like an undirected network, a directed network graph is defined as $G = (V, E)$, but the edges in the network are directed. A directed edge means that two nodes associated with an edge have a certain order relationship. For example, edge $e = (i, j)$ is an ordered pair of nodes i and j, where i is the starting point of e and j is the end point of e. This means, for each edge of the network, there is a definite direction from the beginning to the end of the edge. We use such edges to describe the occurrence of biological reactions. For example, the relationship between the transcription factor and the regulated gene in the regulatory relationship is an orderly relationship. Thus, regulatory networks are often constructed as directed networks. In addition, some computational methods can be used only in directed networks.

Methods based on topological features

Based on network motifs and principal component analysis, Wang *et al.* [28] proposed a new method to characterize the importance of nodes in a directed biological network and applied this method to the following networks: a neural, three

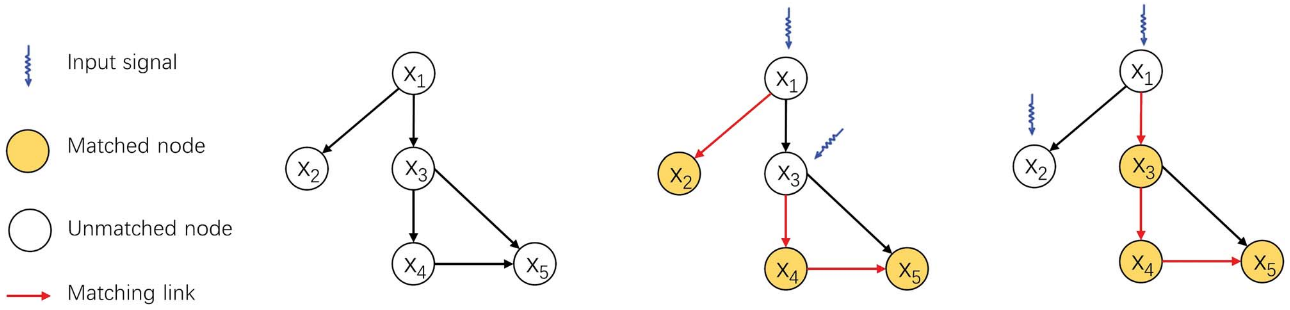

Figure 3. An example of detecting the driver nodes in a directed network. The concept of a maximum matching is illustrated in this figure. Matching is a set of links, where the links do not share start or end nodes. Maximum matching is to find a maximum set of links that do not share start or end nodes. A matched node is the node that is pointed to by a link in maximum matching; otherwise, it is unmatched. The unmatched nodes are driver nodes. Two different results are obtained.

transcriptional regulatory and one signal transduction. Their method performs better than in-degree, out-degree, total degree and other topological methods.

Methods based on system controllability

Liu et al. [45] applied structural controllability to the human signal network represented by a directed graph and detected the driver nodes (Figure 3). Furthermore, they provided a systematic analysis of the role of different proteins in controlling human signal networks. Because fewer driving nodes are needed to control the entire human signal network, the input of different control signals on the regulatory factors that control cancer-related genes may be less expensive than directly controlling cancer-related genes. They introduced a new perspective for the control of human cell signaling systems.

The main challenge in the analysis of system controllability of biological networks is the availability of large-scale biological related networks and an effective tool in analyzing their controllability. In response to this problem, Vinayagam et al. [46] synthesized a directed human PPI network and an analysis framework that describe the structure controllability of a directed weighted network. According to the effect on the number of driver nodes in the network, resulting from the removal of a specific protein, a protein is classified as ‘indispensable’, ‘neutral’ or ‘dispensable’. ‘Indispensable’ nodes are identified as leading disease targets for mutations, viruses and drugs. Experiments show that controllability analysis is very useful in identifying new disease genes and potential drug targets.

The MDS, previously proposed, has been successfully used to analyze large-scale undirected networks and identify cancer-related proteins. Ishitsuka et al. [47] presented an algorithm that uses efficient graph reduction to identify critical control nodes in large-scale directed complex networks, making MDSs available to directed networks. This algorithm is 176 times faster than the existing method and increases the computable network size to 65 000 nodes.

Wu et al. [48] analyzed the controllability of the network structure in the concept of the minimum steering node sets (MSSs). The node started by the input control signal is called a steering node. The MSS consists of a minimum number of nodes driven by input control signals. A comparison of the definitions between MSS and MDS shows that each MSS contains an MDS, whereas the MDS is a maximum subset of the MSS. They developed a new algorithm that treats MSS as a minimum-cost maximum flow problem, can be solved in polynomial time and can find a clear relationship between MDS and MSS. The application results show that MSSs, rather than MDSs, better reflect the dynamics of the network and the importance of nodes for network control. However, because the MSSs of complex networks are not unique, the importance of different MSSs is varied in practical applications. Therefore, MSSs with certain meanings should be studied. Wu et al. [49] proposed a method of studying the MSSs of a biomolecular network by considering drug-binding information. Biomolecules in the MSSs with binding preferences are rich in known drug targets and may have more chemical-binding opportunities with existing drugs than randomly selected MSSs, suggesting that this approach may be drug target identification and drug relocation Promising tools.

Based on recent results for full and structural controllability of directed networks, which enables a set of driver nodes to control the whole network, Kanhaiya et al. [50] presented a new method for the structural controllability of cancer networks and used it to analyze breast, pancreatic and ovarian cancers. They found that although many drugs acting on the driver nodes are part of known cancer therapy, there are some drug target driver nodes identified by the algorithm that are not used for any cancer treatment. Experiments have shown that a better understanding of the dynamics of cancer control through mathematical modeling may pave the way for new effective treatments and personalized medicine.

Ravindran et al. [51] modeled the human cancer signal network as a directed graph and explored its controllability. They used a maximum matching algorithm to identify the driver nodes, which are divided into backbone, peripheral and ordinary according to their role in regulatory interaction and network control. In addition, based on the effect genes have on the size of the driver nodes set, genes are classified as indispensable, dispensable and neutral. Discovering these indispensable backbone nodes is key to driving the regulatory network into a cancer phenotype (via mutations) and to a healthy phenotype (as a drug target).

Noise is widely present in the data of biological networks, which may be detrimental to the accurate analysis of structural controllability. Chu et al. [52] proposed a comprehensive analysis package, WDNfinder, which provides structural controllability analysis using node connection strength in a directed and weighted network. The method includes two innovative algorithms: maximum weight MDS (MWMDS) identification, MMWDS sampling and node classification. Compared to existing methods, WDNfinder can significantly reduce the set of minimum drive node sets (MDS) under the constraints of domain knowledge. This method works for any directed and weighted network.

In weighted network

A weighted network is represented as a weighted graph, $G = (V, E)$. Each edge, $e = (i, j) \in E$, is assigned with a weight, $w_{i,j}$, which represents the relationship between nodes i and j, such as distance, relevance, probability of interaction, etc. Compared with unweighted networks, weighted networks provide the strength of contact between nodes, providing a more accurate analysis of biological research.

Methods based on combination of topological features and biological information

Jiang *et al.* [53] put forward a new computational method, integrated edge weights (IEW), which first ordered the PPIs by integrating their edge weights, then determined the sub PPI network consisting of those highly ranked edges. Finally, the nodes in these subnetworks are considered as essential proteins. They performed IEW assessments on three model organisms: *S. cerevisiae*, *E. coli* and *Caenorhabditis elegans*. Experimental results show that this method performs better than the existing technologies, such as DC, NC, etc.

Li *et al.* [54] studied the topological structure of key proteins from a completely new perspective. This was the first time that a topological potential was used to identify key proteins in a PPI network. The topology potential was first used to describe non-contact interactions between particles of matter. The basic idea is that every protein in the network can be viewed as a particle that creates a potential field around itself and that all protein interactions form a topological field on the network. By defining and calculating the value of the topological potential of each protein, we obtain a more precise ranking that reflects the importance of the protein in the PPI network. Experimental results show that, for predicting key proteins, the topological potential–based methods, TP and TP-NC, are superior to the traditional topological methods. In addition, under the control of topological potential, these central methods have also been improved in identifying the performance of key proteins in biological networks.

At present, most of the methods mainly focus on the topology of PPI networks and largely ignore gene ontology annotation information. Therefore, Zhang *et al.* [55] proposed a new centrality measure, TEO, that identifies essential proteins by combining network topology, gene expression profiles and GO information. To evaluate the performance of the TEO method, they compared it with other methods (degree, mediation, NC, Pec, etc.) in the detection of essential proteins from two different yeast PPI datasets. The results show that adding GO information effectively improves prediction precision. This method is superior to other methods in predicting essential proteins.

To identify essential proteins in the signal transduction network, first, the importance of proteins in the signal transduction network must be understood. However, there are relatively few methods in assessing the importance of proteins in signal networks. Based on the concept of the minimax distance metric algorithm (MDMa), Wang *et al.* [56] developed a new method in assessing the importance of proteins in signal transduction networks. The MDMa in a signal transduction network refers to a set of proteins that propagate signals from the input to the output in a minimal distance. Applying this method to the large-signal transduction network of small cell lung cancer, they found a significant correlation. Useful features that allow the prediction of the criticality and conservation of proteins are observed in signal transduction networks.

Because development costs far exceed sales revenue, the costs, time and risks associated with drug development make the development of rare-disease drugs unattractive for the pharmaceutical industry. The potential strategy to address new drug development challenges and improve existing drug treatment capabilities is to improve the selection of drug targets and study new uses for existing approved drugs. Mazandu *et al.* [57] created a comprehensive computational framework by using drug-target association data sets, as well as the human-pathogen functional interaction network, gene-disease associations, Disease Ontology and GO to improve drug development and repositioning of approved drugs. Their model uses a process-based semantic similarity that performs better than target-based similarity. In addition, this computational framework can be used to examine the applicability of experimentally determined goals and may constitute a useful tool to guide further experiments.

Methods based on system controllability

In the previous method, key proteins were predicted as the high-degree nodes on the PPI network. However, due to lack of data, some key proteins cannot have high connectivity. The network must be redesigned from other biological data sources. Habibi *et al.* [58] defined a minimal set of proteins whose removal would destroy the largest number of protein complexes. Firstly, the author built a weighted graph using a given set of composites. Then, they proposed a more appropriate method based on betweenness values to diagnose a minimal set of proteins described above. The experimental results show that the cut set proposed by the author performs better than other cut sets.

Methods based on machine learning

Most methods ignore the edge dynamics of biomolecular networks under different conditions, using only the 'guilt by association' principle, which limits their performance. To solve this problem, Luo *et al.* [59] proposed an algorithm, DoCE, which combines 'guilt by association' with 'guilt by rewiring' of biomolecular networks for identifying disease genes. First, they weighted the edges of the PPI network by the differences in gene co-expression between the case and control samples to obtain edge dynamics of the biomolecular network. Then, they extracted features from the weighted PPI network. Finally, they used logistic regression to identify disease genes. The algorithm has an area under curve (AUC) value above 0.9. In addition, two new schizophrenia-related genes were discovered.

Zhang *et al.* [60] proposed a new method for identifying complexes on PPI networks based on different co-expressed information. First, they used Markov clustering algorithms and edge-weighted schemes to predict complexes on PPI networks. Then, they proposed some important features such as graphical information and gene expression analysis to filter and modify the predicted complex. The authors evaluated their methods on two yeast PPI networks. Compared with the existing methods, COACH, ClusterONE and MCL, this method increases the precision value by at least 0.1752, the F-Measure value by at least 0.0448 and the Sn value by at least 0.0771.

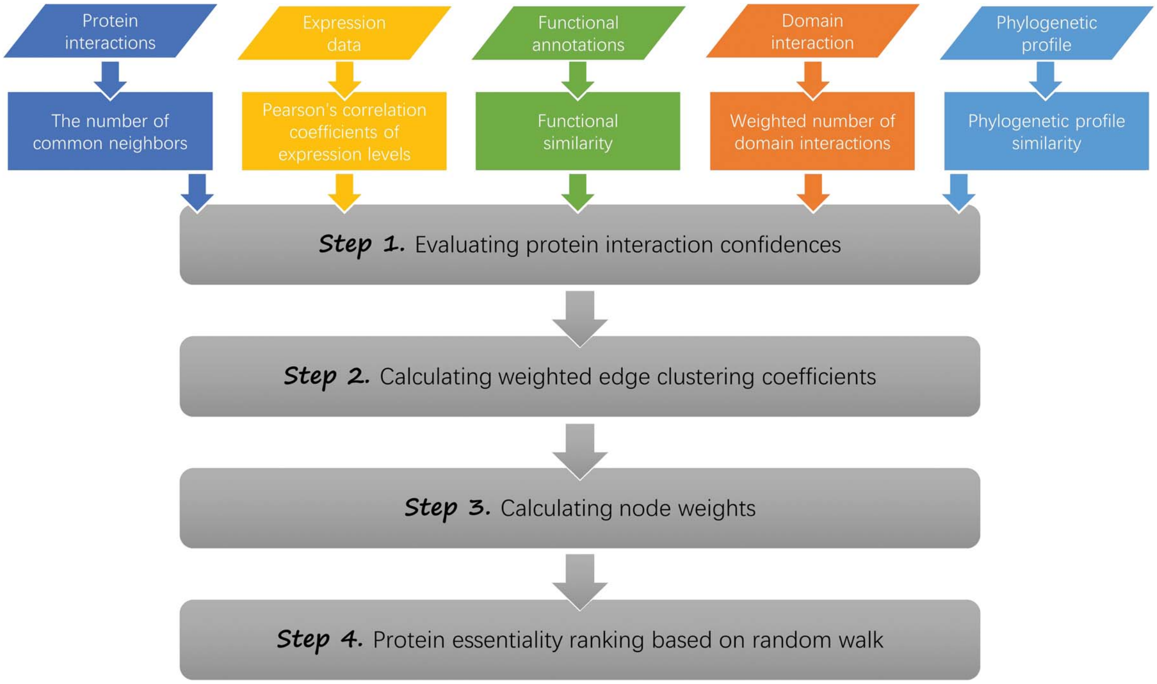
Protein interactions
Expression data
Functional annotations
Domain interaction
Phylogenetic profile
The number of common neighbors
Pearson's correlation coefficients of expression levels
Functional similarity
Weighted number of domain interactions
Phylogenetic profile similarity
Step 1. Evaluating protein interaction confidences
Step 2. Calculating weighted edge clustering coefficients
Step 3. Calculating node weights
Step 4. Protein essentiality ranking based on random walk

Figure 4. The workflow of EssRank.

Methods based on random walk

Can *et al.* [61] proposed random walks to solve complex/path member problems on a graph. Random walk technology explores the global structure of the network by simulating the behavior of random walkers. Based on the probability of connecting edges, a random walker starts from the initial node (query node) and moves to the neighbor node. Finally, the percentage of time spent on the node gives the concept of its proximity to the query node. They evaluated the proposed method in three different probabilistic yeast networks. In addition, they compared the proposed technology with two other existing technologies in terms of accuracy and runtime performance, thus solving the scalability problem of graphical analysis technology for the first time. Experiments show that the random walk method is suitable for predicting candidate members of a core complex or part of a known pathway. Compared with the best competitive technology, random walk technology, at more than 1000 times the speed, achieves similar or higher accuracy.

Similarly, Xu *et al.* [62] introduced a new method, essentiality ranking (EssRank), that ranks the importance of proteins based on random walks to improve the accuracy of essential protein detection. To solve the problem that protein interaction data, obtained by high-throughput experiments, usually contains high false positives, they proposed the following measures: first, five types of biological data were used to evaluate the confidence score of PPIs. Then, a weighted protein interaction network was constructed, where edge weight is calculated according to the confidence score and the weighted edge clustering coefficient (WECC) of the network topology. The weight of the nodes is the sum of the WECC values of its connecting edges. Finally, the importance of protein is calculated iteratively through a personalized PageRank algorithm. Figure 4 shows the workflow of EssRank. Experimental results on the yeast PPI network show that EssRank is superior to most existing methods, including the most commonly used centrality measures (SC, DC, BC, CC and EC), topology-based methods (NC) and the data integration method, IEW.

In dynamic network

Since static interactions in static biological networks are obtained at different moment points and under different conditions, static graphs cannot fully reflect real networks. The dynamic network is the network whose network topology changes with time and the biomolecules in it, which are highly expressed. Therefore, the computational method used in the dynamic network is more accurate and of high-quality performance.

Methods based on combination of topological features and biological information

PPI data obtained from large-scale, high-throughput experiments generally contain incorrect information. Using raw PPI data to identify essential proteins is insufficient. How to improve accuracy has become the focus of identifying essential proteins. Luo *et al.* [63] introduced a new method to predict essential proteins by integrating dynamic LAC and in-degree of proteins in complexes, combine dynamic LAC with complex centrality (CDLC). The CDLC was applied to the PPI network of *S. cerevisiae*. The results show that CDLC outperforms the other methods: DC, LAC, SoECC, etc.

Xiao *et al.* [64] proposed a framework in identifying essential proteins from the active PPI network constructed by dynamic gene expression. First, they filtered the interfering genes based on the dynamic gene expression profile and then constructed an active PPI network. Afterward, they used six typical centrality measures (DC, LAC, NC, BC, CC and SC) to predict essential proteins in the dynamic PPI network. The experimental

Table 1. The summary of the methods mentioned above

Method category	Summary of the methods
Methods based on topological features	The methods of this type only use the topology features of the network. Typical centrality measures (DC, EC, CC, BC and SC) are the simplest but have limited performance. SoECC, NC and LAC combine the relationship between nodes and nodes or the features of edges to improve performance. ME learns topological features from the perspective of network motifs. But these methods can only be used in undirected networks. [28] is a method based on network motifs and multivariate statistical analysis that can identify important nodes in directed biological networks.
Methods based on combination of topological features and biological information	The methods of this type improve their performance by combining biological information such as gene expressions, GO annotation data, orthology, subcellular localization information, etc. CEPPK, CSC, GOS, LID and GEG are some methods of this type used in undirected biological networks. And GEG can improve the prediction when predicting a small number of key proteins. To solve the problem of biological information loss in static networks, some works using biological information to construct weighted or dynamic networks. IEW, TP-NC, TEO and [56, 57] are some methods of this type used in weighted biological networks. CDLC, EPFOA and [64, 66] work in dynamic biological networks.
Methods based on system controllability	The methods of this type consider key nodes identification as the problem of finding the fewest nodes, which can be controlled, to make the network run to the desired target state in a network. CC-MDS, a method used in undirected biological networks, which is able to solve the problem of using different optimization methods will generate different MDSs, based on the theory that a node, which has high degrees and high betweenness, is more likely to be the controller. [45–52] work in directed biological networks. [47] makes MDS available to directed networks. [48] treats MSS as a minimum-cost maximum flow problem, which can be solved in polynomial time. [52] avoids the effects of noise data by adding MWMDS identification, MWMDS sampling and node classification algorithms. [58] builds a weighted network through a given protein complex and solves the problem of the inability to identify an essential protein that cannot have high connectivity due to lack of data.
Methods based on machine learning	The methods of this type extract knowledge from a large quantity of biological data and quickly and accurately recognize critical nodes as node classification problems. But the methods of this type cannot select features automatically and are difficult to select parameters. [42–44] work in undirected biological networks. [42] can predict essential genes when no experimental knock-out data is available. [44] improves classification performance by adding network topological features of the obtained flux-coupled subnetwork. DoCE uses rewired information to build a weighted PPI network to obtain edge dynamics of the biomolecular network. [60] identifies protein complexes based on differential co-expression information based on the idea that different proteins in the same complex have similar trends in gene expression intervals.
Methods based on random walk	The methods of this type provide a new idea in identifying important nodes in biological networks. [61] is suitable in predicting candidate members of core complexes or parts of known pathways. [62] builds a weighted PPI network by combining protein interactions, expression data, functional annotations, domain interaction, phylogenetic profile and network topology and uses a random walk–based approach to sort the importance of proteins to improve the accuracy of key protein detection.

results on the yeast network show that the identification of essential proteins based on the active PPI network significantly improves the performance of the centrality measure in determining the number of essential proteins and recognition accuracy. Likewise, the results also show that most of the key proteins are active.

Lei *et al.* [65] extended the fruit fly optimization algorithm (FOA) to identify essential proteins. This algorithm, EPFOA, combines FOA with topological properties and biological information to identify essential proteins. The algorithm, EPFOA, not completely relying on ranking score identification individually, has the advantage of simultaneously identifying multiple essential proteins. Firstly, the authors combine the gene expression data with the static PPI network to build a dynamic network model. Then they put forward a new topology centrality method based on GO annotation and edge aggregation coefficient (ECC) to measure the LAC of the dynamic PPI network. In addition, the distribution of proteins in each compartment was obtained based on subcellular localization data, and the role of compartments in the identification of essential proteins was analyzed. Finally, EPFOA aims to identify the essential protein. Experimental results show that EPFOA performs better than the existing essential protein detection methods (DC, EC, IC, SC, NC, LAC, PeC, etc.).

A protein can interact with different proteins at different times or conditions. Many existing methods only utilize static PPI network data that may lose a lot of temporal biometric information. To solve this problem, Liu *et al.* [66] combined gene expression data with traditional static PPI networks to construct the dynamic PPI networks. In order to filter out data noise, the authors used GO semantic similarity as the network weight. Finally, after constructing a dynamic PPI network, the authors proposed a method based on 'core-attachment' structure to detect complexes. The experimental results show that

Table 2. The summary of the biological network data sets

Network category	Biological data and sources	Instance [reference number]
PPI network	Protein complex sets: MIPS[1] database. Essential proteins of yeast: MIPS[1], SGD[2], SGDP[3] and DEG[4] databases. PPIs of yeast: DIP[5], MIPS[1], BioGRID[6] databases. Gene expression data of yeast: GEO[7] database. Protein orthologous information: InParanoid[8] database. Protein subcellular localization annotation information of yeast: COMPARTMENTS[9] database. High-quality protein interactions in *H. sapiens*: HINT[10] database. Protein complexes in *H. sapiens*: CORUM[11] database. Disease-associated genes in *H. sapiens*: UniProt[12], OMIM[13] and GAD[14] databases. Human virus-targeted proteins: VirusMINT[15] database. Human transcription factors: TRANSFAC[16] database. PPIs of human: OPHID[17] database. Essential proteins of human: DEG[4], BioMart[18] and DAVID[19] databases. Directed human PPI data: DirectedPPI[20] database. GO and annotation for yeast: Gene Ontology Consortium[21]. Protein domain interactions: DOMINE[22] database.	*Saccharomyces cerevisiae* PPI network [34] *Homo sapiens* PPI network [41] Human PPI network [3] Directed human PPI network [46] *Caenorhabditis elegans* PPI network [53] Dynamic yeast's PPI network [64]
Metabolic network	Genome information of organism-specific metabolic networks: KEGG[23], WIT[24], and Ecocyc[25] databases. Essential gene and non-essential genes in yeast: NMPDR[26] database. Known metabolic reactions and enzymes of Arabidopsis: AraCyc[27] database.	*Escherichia coli* metabolic network [4] *Pseudomonas aeruginosa* metabolic network [42] *Salmonella typhimurium* metabolic network [42] Arabidopsis metabolic network [5]
Gene regulatory network	Transcriptional regulation in *E. coli*: RegulonDB[28] database.	*Caenorhabditis elegans* neural network [9] Yeast transcriptional regulatory network [10] Drosophila developmental transcriptional network [13] Human signal transduction network [11]
Signal transduction network	Mutant phenotype of mouse genes: MGD[29] database. Orthologous genes: Ensembl Gene[30] database. Small cell lung cancer cell signal transduction pathway: KEGG[23] database.	Signaling network in the hippocampal CA1 neuron of a mouse [12] Host immune response network [13] Guard cell ABA signaling network [14] T-cell receptor signaling network [15] Human signaling network [16] An intracellular signal transduction network [7] Human cancer signaling network [53] The large signal transduction network in the small cell lung cancer [56]

Database address:
[a] http://mips.helmholtz-muenchen.de/proj/ppi/
[b] http://www.yeastgenome.org/
[c] http://www-sequence.stanford.edu/group/yeast_deletion_project
[d] http://tubic.tju.edu.cn/deg/
[e] http://dip.doe-mbi.ucla.edu/
[f] http://thebiogrid.org/
[g] http://www.ncbi.nlm.nih.gov/geo/
[h] http://inparanoid.sbc.su.se/cgibin/index.cgi
[i] http://compartments.jensenlab.org
[j] http://hintdb.hgc.jp/hint/
[k] http://mips.helmholtz-muenchen.de/genre/proj/corum/index.html
[l] http://www.uniprot.org/
[m] http://www.omim.org/
[n] http://geneticassociationdb.nih.gov/
[o] http://mint.bio.uniroma2.it/virusmint/
[p] http://www.gene-regulation.com
[q] http://ophid.utoronto.ca/
[r] http://www.biomart.org
[s] http://www.david.niaid.nih.gov
[t] www.flyrnai.org/DirectedPPI
[u] http://www.geneontology.org
[v] http://domine.utdallas.edu
[w] http://www.genomejp/kegg/
[x] http://wit.mcs.anl.gov/WIT2/
[y] https://ecocyc.org/
[z] http://www.nmpdr.org
[aa] http://www.arabidopsis.org/biocyc/
[ab] http://regulondb.ccg.unam.mx/
[ac] http://www.informatics.jax.org/menus/expression_menu.shtml
[ad] http://asia.ensembl.org/index.html

this method performs well in the detection of protein complexes in dynamic weighted PPI networks.

Summary of the methods mentioned

We summarize the methods mentioned above in Table 1.

Biological network data sets

The biological network datasets are summarized in Table 2.

Conclusion

Essential biomolecules, such as essential genes and essential proteins, are related to the survival of organisms and the selection of drug targets. The traditional way of finding essential biomolecules through experimental methods requires extensive manpower, time and money. With the development of technology and the growth of biological information, we have the ability to build many types of biological networks. Researchers model biological networks as follows: undirected, directed, weighted, dynamic, etc. The biological network provides a comprehensive view of the cells, or entire organism, allowing researchers to systematically study a large number of biological collective behaviors and the characteristics of the expression. The identification of essential biomolecules in the organism is regarded as identifying important nodes in the network.

To identify important nodes in biological networks, researchers have proposed a variety of computational methods, based on the following features: topological, a combination of topological features and biological information, machine learning, controllability and random walks. Methods, based on topological features, including a variety of centrality methods, are the simplest, and most of them are suitable for use in undirected and unweighted networks. The method of combining topological features and biological information is based on the topology-based method and takes into account known biological information. Researchers use this method to weight nodes or weight the network and construct weighted networks, dynamic networks, etc. This method is more accurate and reliable than the method based only on topological features. The method based on machine learning extracts knowledge from a large quantity of biological data and quickly and accurately recognizes critical nodes as node classification problems. The method based on system controllability considers key nodes identification as the problem of finding the fewest nodes, which can be controlled, to make the network run to the desired target state in a network. However, these methods all require reliable data. With the development of high-throughput technology, it is possible to obtain thousands of data from biomolecules and intermolecular interactions. For example, Next Generation Sequencing has the advantages of higher throughput, shorter running time, longer sequencing fragments and lower cost than traditional sequencing methods and can generate a large amount of genomic information data at low cost. However, the disadvantage of Next Generation Sequencing is that its polymerase chain reaction (PCR) process increases the error rate of sequencing and has system bias. And sequencing errors have a negative impact on the construction of the network and make it difficult to accurately detect essential genes in the network. Therefore, in the future, to improve the accuracy of the results, researchers are supposed to consider the reliability of data when using these high-throughput data and find ways to reduce the impact of noise data on research, such as the work of Xu *et al.* [62] mentioned above. It is also very meaningful to develop high-throughput technology to produce more reliable data. The development of high-throughput technology promotes the advancement of bioinformatics, and the advancement of bioinformatics facilitates the development of high-throughput technology. In addition, with the rapid growth of biological data, deep learning can be well applied in the biological field. Deep learning methods can extract raw features from massive and annotated data and then explore the laws behind these original features. Compared to the methods based on machine learning, deep learning methods can solve the problem of not being able to select features automatically, and researchers using this type of methods do not need to master a priori knowledge. Deep learning has been applied to many fields in biology, such as works in [67–69], but the application of deep learning methods is rarely seen in the field of identifying important nodes in biological networks. The researchers are supposed to use deep learning to find critical biomolecules in biological networks in future research.

Key Points

- The method of combining topological features and biological information is more accurate and reliable than the method based only on topological features.
- The method based on controllability identifies key nodes by looking for driver nodes. The method based on machine learning recognizes critical nodes as node classification problems. These two kinds of methods provide new ideas for key nodes identification.
- The more complete the network (such as directed, weighted, dynamic network), the better the performance of the computational method.

Funding

The work was supported by the National Natural Science Foundation of China (61472333, 61772441, 61472335, 61272152 and 41476118); Project of marine economic innovation and development in Xiamen (16PFW034SF02); Natural Science Foundation of the Higher Education Institutions of Fujian Province (JZ160400); Natural Science Foundation of Fujian Province (2017J01099); President Fund of Xiamen University (20720170054); and Juan de la Cierva position (code: IJCI-2015-26991 to X.Z.).

References

1. Liu J-G, Lin J-H, Guo Q, *et al*. Locating influential nodes via dynamics-sensitive centrality. *Sci Rep* 2016;**6**:032812.
2. Kim W, Li M, Wang J, *et al*. Essential protein discovery based on network motif and Gene Ontology. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2011*, pp. 470–5. IEEE, Atlanta, GA.
3. Le D-H, Xuan Hoai N, Kwon Y-K. A comparative study of classification-based machine learning methods for novel disease gene prediction. *Knowledge and Systems Engineering*. 2015, 577–88.

4. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 2003;**19**:1423–30.
5. Clark ST, Verwoerd WS. A systems approach to identifying correlated gene targets for the loss of colour pigmentation in plants. *BMC Bioinform* 2011;**12**:343.
6. Li J, Zhao PX. Mining functional modules in heterogeneous biological networks using multiplex PageRank approach. *Front Plant Sci* 2016;**7**:903.
7. Helikar T, Konvalina J, Heidel J, *et al*. Emergent decision-making in biological signal transduction networks. *Proc Natl Acad Sci U S A* 2008;**105**:1913–8.
8. Ravindran V, Sunitha V, Bagler G. Controllability of human cancer signaling network. In: *2016 International Conference on Signal Processing and Communication*, *2016*, pp. 363–7. IEEE, Noida, India.
9. Varshney LR, Chen BL, Paniagua E, *et al*. Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Comput Biol* 2011;**7**:e1001066.
10. Costanzo MC, Crawford ME, Hirschman JE, *et al*. YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information. *Nucleic Acids Res* 2001;**29**:75–9.
11. Milo R, Itzkovitz S, Kashtan N, *et al*. Superfamilies of evolved and designed networks. *Science* 2004;**303**:1538–42.
12. Jenkins SL, Neves S, Hasseldine A, *et al*. Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science* 2005;**309**:1078–83.
13. Thakar J, Pilione M, Kirimanjeswara G, *et al*. Modeling systems-level regulation of host immune responses. *PLoS Comput Biol* 2007;**3**:e109.
14. Li S, Assmann SM, Albert R. Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biol* 2006;**4**:e312.
15. Saezrodriguez J, Simeoni L, Lindquist JA, *et al*. A logical model provides insights into T cell receptor signaling. *PLoS Comput Biol* 2007;**3**:e163.
16. Cui Q, Yun M, Jaramillo M, *et al*. A map of human cancer signaling. *Mol Syst Biol* 2014;**3**:152.
17. Ren J, Wang J, Li M, *et al*. Prediction of essential proteins by integration of PPI network topology and protein complexes information. In: *International Symposium on Bioinformatics Research and Applications, 2011*, pp. 12–24. Springer, Berlin, Heidelberg.
18. Li M, Zhang H, Wang JX, *et al*. A new essential protein discovery method based on the integration of protein–protein interaction and gene expression data. *BMC Syst Biol* 2012;**6**:15.
19. Zhang X, Xu J, Xiao WX. A new method for the discovery of essential proteins. *PLoS One* 2013;**8**:e58763.
20. Liu W, Li D, Zhang J, *et al*. SigFlux: a novel network feature to evaluate the importance of proteins in signal transduction networks. *BMC Bioinformatics* 2006;**7**:515.
21. Wang RS, Albert R. Elementary signaling modes predict the essentiality of signal transduction network components. *BMC Syst Biol* 2011;**5**:44.
22. Li M, Wang J, Wang H, *et al*. Essential proteins discovery from weighted protein interaction networks. In: *International Conference on Bioinformatics Research and Applications*, *2014*, pp. 89–100. Springer, Berlin, Heidelberg.
23. Peng W, Wang J, Wang W, *et al*. Iteration method for predicting essential proteins based on orthology and protein–protein interaction networks. *BMC Syst Biol* 2012; **6**:1–17.
24. Proulx SR, Promislow DE, Phillips PC. Network thinking in ecology and evolution. *Trends Ecol Evol* 2005;**20**:345–53.
25. Junker BH, Schreiber F. *Analysis of Biological Networks*. Hoboken, New Jersey: John Wiley & Sons, 2011.
26. Seringhaus M, Paccanaro A, Borneman A, *et al*. Predicting essential genes in fungal genomes. *Genome Res* 2006;**16**: 1126–35.
27. Li Y, Wang M, Wang H, *et al*. Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci Rep* 2014;**4**:5765.
28. Wang P, Lü J, Yu X. Identification of important nodes in directed biological networks: a network motif approach. *PLoS One* 2014;**9**:e106132.
29. Koschützki D, Schreiber F. Comparison of centralities for biological networks. In: *Proceedings of the German Conference on Bioinformatics (GCB'04)*, *2004*, pp. 199–206. German Conference on Bioinformatics, Bielefeld, German.
30. Joy MP, Brock A, Ingber DE, *et al*. High-betweenness proteins in the yeast protein interaction network. *BioMed Res Int* 2005;**2005**:96–103.
31. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 2006;**6**:35–40.
32. Wang H, Li M, Wang J, *et al*. A new method for identifying essential proteins based on edge clustering coefficient. *Lecture Notes in Computer Science*, Vol. 6674. 2011, 87–98.
33. Wang J, Li M, Wang H, *et al*. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**:1070–80.
34. Li M, Wang J, Chen X, *et al*. A local average connectivity-based method for identifying essential proteins from the network level. *Comput Biol Chem* 2011;**35**:143–50.
35. Li G, Zhang X, Luo J, *et al*. A network motif extension-based method for identifying essential proteins from protein–protein interaction networks. *J Comput Theor Nanosci* 2017;**14**:1682–9.
36. Li M, Zheng R, Zhang H, *et al*. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. *Methods* 2014;**67**:325–33.
37. Luo J, Ma L. A new integration-centric algorithm of identifying essential proteins based on topology structure of protein–protein interaction network and complex information. *Curr Bioinform* 2013;**8**:380–5.
38. Li M, Niu Z, Chen X, *et al*. A reliable neighbor-based method for identifying essential proteins by integrating gene expressions, orthology,and subcellular localization information. *Tsinghua Sci Technol* 2016;**21**:668–77.
39. Qi Y, Luo J. Prediction of essential proteins based on local interaction density. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**13**:1170–82.
40. Zhang W, Xu J, Li X, *et al*. A new method for identifying essential proteins by measuring co-expression and functional similarity. *IEEE Trans Nanobioscience* 2016;**15**: 939–45.
41. Zhang X-F, Ou-Yang L, Zhu Y, *et al*. Determining minimum set of driver nodes in protein–protein interaction networks. *BMC Bioinformatics* 2015;**16**:146.
42. Plaimas K, Eils R, König R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol* 2010;**4**:56.
43. Hor C-Y, Yang C-B, Yang Z-J, *et al*. Prediction of protein essentiality by the support vector machine with statistical tests. *Evol Bioinform* 2013;**9**:387–416.

44. Nandi S, Subramanian A, Sarkar RR. An integrative machine learning strategy for improved prediction of essential genes in *Escherichia coli* metabolism using flux-coupled features. *Mol BioSyst* 2017;**13**:1584–96.
45. Liu X, Pan L. Identifying driver nodes in the human signaling network using structural controllability analysis. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:467–72.
46. Vinayagam A, Gibson TE, Lee HJ, *et al*. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc Natl Acad Sci U S A* 2016;**113**:4976–81.
47. Ishitsuka M, Akutsu T, Nacher JC. Critical controllability analysis of directed biological networks using efficient graph reduction. *Sci Rep* 2017;**7**:14361.
48. Wu L, Li M, Wang J, *et al*. Minimum steering node set of complex networks and its applications to biomolecular networks. *IET Syst Biol* 2016;**10**:116–23.
49. Wu L, Tang L, Li M, *et al*. Biomolecular network controllability with drug binding information. *IEEE Trans Nanobioscience* 2017;**16**:326–32.
50. Kanhaiya K, Czeizler E, Gratie C, *et al*. Controlling directed protein interaction networks in cancer. *Sci Rep* 2017;**7**:10327.
51. Ravindran V, Sunitha V, Bagler G. Identification of critical regulatory genes in cancer signaling network using controllability analysis. *Phys A* 2017;**474**:134–43.
52. Chu Y, Wang Z, Wang R, *et al*. WDNfinder: a method for minimum driver node set detection and analysis in directed and weighted biological network. *J Bioinform Comput Biol* 2017;**15**:1750021.
53. Jiang Y, Wang Y, Pang W, *et al*. Essential protein identification based on essential protein–protein interaction prediction by Integrated Edge Weights. *Methods* 2015;**83**:51–62.
54. Li M, Lu Y, Wang J, *et al*. A topology potential-based method for identifying essential proteins from PPI Networks. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:372–83.
55. Zhang W, Xu J, Li Y, *et al*. Detecting essential proteins based on network topology, gene expression data, and Gene Ontology information. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**15**:109–16.
56. Wang T, Xue J, Tan W, *et al*. Learning and identifying the crucial proteins in signal transduction networks by a novel method. In: *International Conference on Computer Science and Education*, *2014*, pp. 15–9. IEEE, Vancouver, BC, Canada.
57. Mazandu GK, Chimusa ER, Rutherford K, *et al*. Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. *Brief Bioinform* 2017;**19**(6):1141–52.
58. Habibi M, Khosravi P. Disruption of the protein complexes from weighted complex networks. *IEEE/ACM Trans Comput Biol Bioinform* 2018:1–1.
59. Luo P, Tian LP, Ruan J, *et al*. Identifying disease genes from PPI networks weighted by gene expression under different conditions. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2016*, pp. 1259–64. IEEE, Shenzhen, China.
60. Zhang Z, Song J, Tang J, *et al*. Detecting complexes from edge-weighted PPI networks via genes expression analysis. *BMC Syst Biol* 2018;**12**:40.
61. Can T, Çamoĝlu O, Singh AK. Analysis of protein–protein interaction networks using random walks. In: *Proceedings of the 5th International Workshop on Bioinformatics, 2005*, pp. 61–8. ACM, Chicago, Illinois.
62. Xu B, Guan J, Wang Y, *et al*. Essential protein detection by random walk on weighted protein–protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2017: 1–1.
63. Luo J, Kuang L. A new method for predicting essential proteins based on dynamic network topology and complex information. *Comput Biol Chem* 2014;**52**:34–42.
64. Xiao Q, Wang J, Peng X, *et al*. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. *BMC Genomics* 2015;**16**:Suppl 3:S1.
65. Lei X, Ding Y, Fujita H, *et al*. Identification of dynamic protein complexes based on fruit fly optimization algorithm. *Knowl Based Syst* 2016;**105**:270–7.
66. Liu L, Sun X, Song W, *et al*. A method for predicting protein complexes from dynamic weighted protein–protein interaction networks. *J Comput Biol* 2018;**25**:586–605.
67. Habibi M, Weber L, Neves M, *et al*. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 2017;**33**:i37–48.
68. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**:851–69.
69. Sun T, Zhou B, Lai L, *et al*. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform* 2017;**18**:277.