Briefings in Bioinformatics, 00(00), 2019, 1–18

doi: 10.1093/bib/bbz017
Advance Access Publication Date: 18 February 2019
Review article

Network-based methods for predicting essential genes or proteins: a survey

Xingyi Li, Wenkai Li, Min Zeng, Ruiqing Zheng and Min Li

Corresponding author: Min Li, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China; Tel.: +86-731-88830212; Fax: +86-731-88830212; Email: limin@mail.csu.edu.cn.

Abstract

Genes that are thought to be critical for the survival of organisms or cells are called essential genes. The prediction of essential genes and their products (essential proteins) is of great value in exploring the mechanism of complex diseases, the study of the minimal required genome for living cells and the development of new drug targets. As laboratory methods are often complicated, costly and time-consuming, a great many of computational methods have been proposed to identify essential genes/proteins from the perspective of the network level with the in-depth understanding of network biology and the rapid development of biotechnologies. Through analyzing the topological characteristics of essential genes/proteins in protein–protein interaction networks (PINs), integrating biological information and considering the dynamic features of PINs, network-based methods have been proved to be effective in the identification of essential genes/proteins. In this paper, we survey the advanced methods for network-based prediction of essential genes/proteins and present the challenges and directions for future research.

Key words: essential genes/proteins; network-based methods; topological characteristics; biological information; dynamic features

Introduction

Essential genes and proteins are closely related to the cell metabolism, differentiation and apoptosis [1]. The identification of essential genes or proteins is mainly for the following reasons. From a theoretical perspective, it contributes to know the minimum demands for cell survival well [2, 3], which plays a vital role in synthetic biology. From a practical perspective, as essential genes or proteins are indispensable for bacterial survival, they are also used as drug targets [4] of new antibiotics. Furthermore, some studies have shown that essential genes are closely related to pathogenic genes [5], thus the prediction of essential genes is of great significance for the discovery of pathogenic genes [6, 7].

Xingyi Li received the BS degree in communication engineering from Central South University, China, in 2015. Currently, she is working toward the PhD degree in the School of Computer Science and Engineering, Central South University, Changsha, China. Her current research interests include bioinformatics and systems biology.

Wenkai Li received the BS degree in computer science and technology from Henan University, China, in 2015, and the MS degree in computer science from Central South University in 2018. His current research interests include bioinformatics and systems biology.

Min Zeng received the BS degree from Lanzhou University in 2013 and the MS degree from Central South University in 2016. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His research interests include bioinformatics, machine learning and deep learning.

Ruiqing Zheng received the BS and the MS degrees in computer science from Central South University, Changsha, China, in 2013 and 2016, respectively. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His research interests include bioinformatics and systems biology.

Min Li received the BS degree in communication engineering and the MS and PhD degrees in computer science from Central South University, Changsha, China, in 2001, 2004 and 2008, respectively. She is currently the vice dean and a professor at the School of Computer Science and Engineering, Central South University. Her main research interests include bioinformatics and systems biology.

Submitted: 31 October 2018; **Received (in revised form):** 21 January 2019

In the biological field, the prediction of essential genes mainly relies on laboratory strategies, such as gene knockout [8, 9], RNA interference [10, 11], antisense RNA [12] and transposon mutagenesis [13, 14]. However, as the number of genes in cells is enormous, experimental techniques are undoubtedly complicated. Moreover, the time of experimental techniques is overhead and these techniques hardly apply to some complex organisms, especially humans. Thus, some researchers began to develop computational methods to identify essential genes/proteins.

In organisms, genes or proteins do not function independently; the interactions between genes or proteins exist extensively and maintain the stability of internal environment [15, 16]. Many of the interactions between genes/proteins have been identified by experimental techniques such as yeast two-hybrid [17], tandem affinity purification and mass spectrometry [18] or by computational methods [19, 20]. The complex biological networks consisting of the interacted genes/proteins are generally scale-free [21], meaning that there are only a few highly connected nodes and many rarely connected nodes in the networks. In the scale-free networks, removing highly connected nodes will be more possibly to destroy the connectivity of networks or increase the shortest path length between nodes [22]. The previous studies have proved that highly connected nodes are more essential than rarely connected nodes in maintaining the functions of scale-free biological networks or the stability of organisms [23–25], which has been verified in the yeast, nematode and fly [26–28]. This means that the essentiality of genes or proteins is possible to be predicted through analyzing the topological characteristics of nodes in the biological networks. Up to now, various topology-based methods have been proposed to predict essential genes or proteins [29–31].

However, there are still certain constraints on the topology-based analysis. On the one hand, a significant proportion of protein-protein interaction (PPI) data is incomplete and contains a lot of noises, which may lead to the inaccurate identification of essential genes or proteins only based on topological methods. On the other hand, the intrinsic biological properties of genes or proteins in the organisms are not fully considered. Based on the biological characteristics of essential genes or proteins, such as conservation, modularity and dynamics, integrating biological information can help further study the functional relevance among genes/proteins. Thus methods integrating protein–protein interaction networks (PINs) and biological information, dynamic network-based methods and machine learning-based methods have been proposed to predict essential genes or proteins [32–34].

In this review, our purpose is to summarize the works related to the network-based identification of essential genes or proteins and attempts to help readers have acquaintance with the current developments and further directions in this field. The paper is organized from the following aspects. First, available databases related to essential genes/proteins are introduced. In Topology-based methods to predict essential genes or proteins, we review the topological methods that characterize the gene/protein essentiality. In Integrating PINs and biological information to predict essential genes or proteins, we present an overview of methods that integrate network topological and biological properties. In Dynamic network-based methods to predict essential genes or proteins, the dynamic network-based prediction methods are presented. In Machine learning-based methods to predict essential genes or proteins, we introduce the prediction methods based on machine learning. In Tools and applications, we introduce the tools and applications related to the prediction of essential genes or proteins based on network. To the end, challenges and future work are discussed.

Essential genes/proteins and related information

Essential genes or proteins

At present, the known essential genes or proteins mainly come from the following databases: DEG [35], MIPS [36], SGD [37], SGDP, OGEE [38], and EGGS [39]. Researchers can harness these types of data to study intrinsic characteristics of essential genes/proteins, reveal the characteristics that are tightly bound to the essentiality and finally propose computational methods to predict essential genes/proteins. Meanwhile, these data, which is verified by experiments, can be used as the golden standard data to verify the prediction results. Moreover, there is a database for collecting some predicted essential genes: pDEG [40]. A brief description of these essential gene databases is listed in Table 1.

Related biological information

The biological information closely related to essential genes or proteins is based on topological characteristics in PINs and biological information such as microarray data, RNA-seq data, protein domains, orthology, subcellular localization, gene ontology and protein complex. A brief description of the databases of related biological information is listed in Table 2.

Protein–protein interaction networks

Proteins are not independent, but interact with each other to maintain the stability of the internal environment of cells. By analyzing the PINs, comprehensive information can be obtained and intricate relationships that manage cellular activities can be revealed. It has been shown that the essentiality of a protein is closely associated with its network centralities in PINs.

There are a large number of publicly available PPI databases, such as DIP [41], STRING [42], BioGRID [43], HPRD [44], MIPS [45], MINT [46], IntAct [47]. Each database has its own property, including a large number of data types and different levels of annotation; PINs can be constructed by using one database or integrated from multiple databases based on certain properties. Due to the PPI data from multi-sources, reliable PINs can be obtained by weighting edges or filtering out the false-positive edges. Meanwhile, some databases, such as STRING, IntAct and MINT, also provide scores with reliability of interactions obtained from different sources. Researchers can filter out interactions with low scores by setting thresholds to obtain more reliable PPI data.

Microarray and RNA-seq data

Microarray data can be used to analyze the changes of molecular abundance, the correlation between genes and the activities of genes under different conditions. RNA-seq data is used to determine the sequence, structure and abundance of RNA molecules in a specific sample; it can provide a more comprehensive gene expression profile. More precise results are expected to be obtained through the use of RNA-seq data rather than microarray data. Gene Expression Omnibus (GEO) is the largest and publicly available database that stores microarray and RNA-seq data. The microarray and RNA-seq data are generally used together with PINs.

Protein domains

Proteins are usually made up of one or more functional domains and different combinations of domains generate a wide variety

Table 1. Databases of essential genes or proteins

Database	Scale	Description	URL
DEG	44 bacteria 1 archaea 8 eukaryotes	Database of Essential Genes. A database of currently available essential genomic elements.	http://www.essentialgene.org
pDEG	16 mycoplasma genomes	Database for storing predicted essential genes.	http://tubic.org/pdeg/
MIPS	Saccharomyces cerevisiae	The Munich Information Center for Protein Sequences. A data resource maintaining the yeast genome database that can extract essential genes in yeast genome.	http://mips.gsf.de
SGD	Saccharomyces cerevisiae	Saccharomyces Genome Database. A database storing information about budding yeast DNA and protein sequences, genetics, cell biology and the associated community of researchers.	https://www.yeastgenome.org
SGDP	Saccharomyces cerevisiae	Saccharomyces Genome Deletion Project. A web page containing the essential and non-essential open reading frames in the *S. cerevisiae* genome.	http://www-sequence.stanford.edu/group/yeast_deletion_project
OGEE	9 eukaryotes 39 prokaryotes	Online GEne Essentiality database. A database storing experimentally verified essential and non-essential genes as well as related gene features.	http://ogee.medgenius.info/browse
EGGs	Bacteria	Essential Genes on Genome Scale. A database containing microbial gene essentiality data.	http://www.nmpdr.org/FIG/eggs.cgi

of proteins. Accordingly, identifying the domains that occur within proteins helps to understand their functions. Studies have shown that the essentiality of a protein may be preserved through the function of protein domains or the combination of domains, rather than entire proteins [48]. The data of protein domains can be downloaded from the Pfam database, which is a repository of protein families [49].

Orthology information

Orthologs are homologous proteins that come from a common ancestor in the process of biological evolution. They usually have high sequence similarities and have the same or extremely similar functions. The information of orthologous proteins can be collected in the databases of Clusters of Orthologous Groups of proteins (COG) [50], OrthoMCL [51], Orthologous Matrix (OMA) [52], IsoBase [53], Inparanoid [54], OrthoDB [55] and EggNOG [56].

Subcellular localization information

Proteins should be located in suitable subcellular compartments to perform their functions and PPIs only occur when proteins are in the same subcellular compartment. At present, the subcellular localization information is mainly used to filter false-positive edges in PINs, that is, physical interactions do not occur when the two proteins are not in the same subcellular localization. The subcellular localizations in a eukaryotic cell are usually divided into the following 11 categories: cytoskeleton, cytosol, endoplasmic, endosome, extra-cellular, Golgi, mitochondrion, nucleus, peroxisome, plasma and vacuole. Currently, in addition to some comprehensive protein localization databases such as UniProtKB, there are also some specific protein localization databases. For instance, FunSecKB2 [57] for fungal proteins, PlantSecKB [58] for plant proteins, MetazSecKB [59] for human and animal, LocDB [60] for human and Arabidopsis and COMPARTMENTS [61], which contains the subcellular localization information of several species.

Gene ontology

The Gene ontology (GO) defines concepts used to represent gene functions and relationships between these concepts. GO can be divided into the following three aspects: cell components, molecular functions and biological processes. A directed acyclic graph can represent these three ontologies, where nodes represent terms, edges correspond to their relationships. Ontology data can be downloaded from the Gene Ontology database [62, 63].

Protein complex

Protein complex is a collection of proteins that forms a multimolecular mechanism in the same time and space [64, 65]; they are the basis of many life activities and can accomplish quite a lot of biological functions. The standard protein complex data can be downloaded from the CYC2008 [66] and MIPS/CORUM [45, 67] databases.

Topology-based methods to predict essential genes or proteins

Removing some genes or proteins will bring huge changes into the network, such as causing PINs to rapidly collapse into isolated nodes or clusters [23], which will break the function of the modules or the interactions in key biological processes. Therefore, the essentiality of a gene or protein is closely related to its topological characteristics in PINs. Topology-based methods score genes or proteins by their centralities in the PINs and their sorting scores are subsequently used to determine if the genes or proteins are essential.

A PIN can be denoted as a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes corresponding to genes, metabolites or proteins. $E = \{e_1, e_2, \ldots, e_m\}$ is the set of edges corresponding to the connections among them, n and m are the number of nodes and edges, respectively. The adjacency matrix of a graph is recorded as $A = (a_{ij})$, and $a_{ij} = 1$ if and only if there is a connection between the node v_i and v_j, otherwise $a_{ij} = 0$.

Table 2. Databases of related biological information

Information	Databases	Scale	Description	URL
PPIs	DIP	834 organisms 28 826 genes/proteins 81 762 interactions	Database of Interacting Proteins. It specifically stores experimentally verified PPIs from literature reports.	http://dip.mbi.ucla.edu/dip
	STRING	2031 organisms 9 643 763 genes/proteins 1 380 838 440 interactions	Search Tool for the Retrieval of Interacting Genes/Proteins. A database that includes proteins and their functional interactions.	https://string-db.org
	BioGRID	15 892 753 genes/proteins 188 517 interactions	Biological General Repository for Interaction Datasets. A database of repository for physical and genetic interactions.	https://thebiogrid.org
	HPRD	30 047 genes/proteins 41 327 interactions	Human Protein Reference Database. A comprehensive database containing various information such as protein annotation, PPIs, post-transcriptional modification, and subcellular localization.	http://www.hprd.org
	MIPS-MPPI	992 genes/proteins 937 interactions	MIPS Mammalian Protein–Protein Interaction Database. A database of high-quality mammalian PPIs obtained from literature mining techniques.	http://mips.helmholtz-muenchen.de/proj/ppi
	MINT	648 organisms 25 530 genes/proteins 125 464 interactions	The Molecular INTeraction Database. A database for storing experimentally validated molecular interactions which are extracted from literatures.	https://mint.bio.uniroma2.it
	IntAct	99 807 genes/proteins 751 286 interactions	IntAct Molecular Interaction Database. A database providing the data of biomolecular interactions and corresponding analysis tools.	https://www.ebi.ac.uk/intact
Gene expression data	GEO	4348 data sets 99 667 series 18 654 platforms 2 537 944 samples	Gene Expression Omnibus. The largest and publicly available database. It includes array-based and sequence-based data.	https://www.ncbi.nlm.nih.gov/geo
Protein domains	Pfam	16 712 families 604 clans	The protein families database. A database of protein families.	http://pfam.xfam.org
Orthology information	COG	4623 clusters of orthologous groups	Clusters of Orthologous Groups. A database derived from extensive comparisons of complete prokaryotic protein sequences.	https://www.ncbi.nlm.nih.gov/COG
	OrthoMCL	150 genomes 1 398 546 protein sequences 124 740 ortholog groups	A database for constructing orthologous groups using a Markov Cluster method.	http://orthomcl.org
	OMA	1617 bacteria 141 archaea 327 eukaryota	Orthologous Matrix. A database for the inference of orthologs among integrated genomes.	https://omabrowser.org
	IsoBase	5 eukaryotic species 87 773 protein sequences 12 693 orthologs 48 120 constituent proteins	IsoRank PPI networks Alignment Based Ortholog Database. A database of functionally related orthologs.	http://isobase.csail.mit.edu
	Inparanoid	273 organisms 3 718 323 sequences	A database providing the orthologs obtained by the InParanoid algorithm.	http://InParanoid.sbc.su.se
	EggNOG	2031 organisms 190 000 orthologous groups	Evolutionary genealogy of genes: nonsupervised Orthologous Groups. A database of functional descriptions and annotations for orthologous groups.	http://eggnogdb.embl.de
	OrthoDB	659 eukaryota 3663 bacteria 345 archaea 3139 viruses	The Hierarchical Catalog of Orthologs. A database combining GO and InterPro to describe orthologs.	https://www.orthodb.org

Continued

Table 2. (continued)

Information	Databases	Scale	Description	URL
Subcellular localization information	FunSecKB2	1 976 832 proteins	A fungal protein subcellular location knowledgebase.	http://bioinformatics.ysu.edu/secretomes/fungi2/index.php
	PlantSecKB	1 415 921 proteins	The plant secretome and subcellular proteome knowledgebase.	http://bioinformatics.ysu.edu/secretomes/plant/index.php
	MetazSecKB	4 080 818 proteins	The human and animal secretome and subcellular proteome knowledgebase.	http://bioinformatics.ysu.edu/secretomes/animal/index.php
	LocDB	total number of proteins: 13 342 human 6262 *Arabidopsis thaliana*	Protein Localization Database for Human and Arabidopsis.	https://www.rostlab.org/services/locDB
	COMPARTMENTS	20 021 by knowledge 4841 by experiments. 1468 by text mining 15 788 by sequence-based predictions	A comprehensive database of subcellular location of proteins.	http://compartments.jensenlab.org
Gene ontology	Gene Ontology database	Over 600 000 experimentally-supported GO annotations	A comprehensive database of functions of genes and proteins.	http://www.geneontology.org
Protein complex	CYC2008	408 heteromeric protein complexes	A comprehensive database of 408 manually curated heteromeric protein complexes that are reliably supported by small-scale experiments reported in the current literature.	http://wodaklab.org/cyc2008
	MIPS-CORUM	7570 protein complexes	The comprehensive resource of mammalian protein complexes.	http://mips.gsf.de/genre/proj/corum/index.html

Generally, the topology-based methods can be divided into the following four types: neighborhood-based methods, path-based methods, eigenvector-based methods and methods that combine multiple topological centralities. Table 3 lists the summary of topology-based methods and a toy network is given to show the computation of classical topology-based methods (Figure 1).

Neighborhood-based methods

Neighborhood-based methods estimate the essentiality of a node by considering its neighbors. Typical neighborhood-based measures are degree centrality (DC) [23], local average connectivity (LAC) [29], edge-clustering coefficient centrality (NC) [30], density of maximum neighborhood component (DMNC) [68] and the topology potential-based (TP) centrality [69].

Degree centrality

DC [23], the well-known and simplest network-based method, holds that the more neighbors of a node has, the great it impacts. It is noteworthy that nodes with same degree have different influences in networks of different sizes. For comparison, the normalized DC value of a given node v_i is defined as the following equation.

$$DC(i) = \frac{k_i}{n-1}, \tag{1}$$

where $k_i = \sum_{j \in N_i} e_{i,j} \mid e_{i,j} \in E$ and N_i corresponds to the neighbor set of v_i.

The DC index has the characteristics of simplicity, intuition and low computational complexity. In some studies, such as network robustness and vulnerability, degree-targeted attack is more effective for scale-free or exponential networks compared with betweenness centrality (BC), closeness centrality (CC) and eigenvector centrality (EC). The disadvantage of DC is that only the information that has the most direct influence on nodes is considered, and the information such as higher-order neighbors is not discussed in more detail, thus it is not accurate enough in many cases.

Local average connectivity

Although essential proteins usually have high connectivity, a certain number of highly connected proteins are not essential and few neighbors of these proteins have been found to interact with each other. In order to distinguish between essential and nonessential proteins with high connectivity, the method LAC [29] is proposed. For a given node v_i, its LAC value is defined as the following equation.

$$LAC(i) = \frac{\sum_{w \in S_i} DC^{C_i}(w)}{|S_i|}, \tag{2}$$

where S_i is the set of neighbors of v_i. C_i is a subnetwork consisting of nodes in S_i and edges between these nodes. For a node w in S_i, its local connectivity $DC^{C_i}(w)$ is defined as how many other nodes in C_i it connects directly.

Edge-clustering coefficient centrality

Traditional centrality methods measure the importance of nodes in PPI networks, but often ignore the importance of the edges between nodes. Some researches indicate that there is a close relationship between nodes and edges in PPI networks. NC [30] is

Table 3. Summary of network-based methods

Methods	Short Description	Network	Data	Reference
DC	DC calculates the number of neighbors in the network.	Unweighted	PPIs	[23]
LAC	LAC measures the importance of a node based on the local average connectivity.	Unweighted	PPIs	[29]
NC	NC scores a node by calculating the edge-clustering coefficient between the node and its neighbors.	Unweighted	PPIs	[30]
DMNC	DMNC calculates the density among the neighbors of this node.	Unweighted	PPIs	[68]
TP	TP calculates the topology potential of each node.	Weighted Unweighted	PPIs	[69]
CC	CC calculates the average of distance between the node and all other nodes in the network to eliminate the interference of special values.	Unweighted	PPIs	[74]
IC	IC measures the importance of nodes by the amount of information that is propagated in paths.	Unweighted	PPIs	[75]
BC	BC measures the number of shortest paths passing through the node.	Unweighted	PPIs	[76]
SC	SC calculates the number of closed loops in which the node appears.	Unweighted	PPIs	[77]
EC	EC calculates the essentiality of a node depends on both the number of its neighbors and the importance of each neighbor.	Unweighted	PPIs	[81]
PR	PR measures the importance of a node depends on the quantity and quality of the other nodes pointing to it.	Unweighted	PPIs	[82]
LR	LR adds a ground node who connects to all nodes through directional links to replace the parameter d in PR.	Unweighted	PPIs	[83]
HITs	HITs uses different indicators to score nodes in the network.	Unweighted	PPIs	[84]
GOS	GOS purifies PPI networks and uses a random walk model to integrate a protein's topological features and its orthology.	Unweighted	PPIs; Gene expression profiles; Subcellular localization information; Orthology information	[94]
SPP	SPP partitions PPI networks by subcellular localization data to obtain subnetworks with different priorities.	Unweighted	PPIs; Subcellular localization information;	[96]
ION	ION integrates the orthology with PPI networks and uses an iteration method to obtain the rank of proteins.	Weighted	PPIs; Orthology information	[98]
UDoNC	UDoNC calculates the number and the frequency of protein domain types and the edge clustering coefficient to score proteins.	Weighted	PPIs; Protein domains	[101]
RSG	RSG constructes a weighted PPI networks by the GO annotation and Pearson correlation of RNA-seq data, and the weighted edge clustering coefficient is used to measure the connectivity of nodes.	Weighted	PPIs; Gene expression profiles (RNA seq and microarray); Subcellular localization information; GO information	[32]
LIDC	LIDC integrates a local interaction density of PPI networks with protein complex information by an integration strategy for multiple bioinformatics.	Unweighted	PPIs; Protein complex	[105]
SCP	SCP integrates the subcellular compartments, Pearson correlation of gene expression and PPI networks.	Weighted	PPIs; Gene expression profiles	[99]
PEMC	PEMC integrates modularity and conservatism of proteins in the PPI networks.	Weighted	PPIs; Gene expression profiles; Protein domains; Orthology information	[100]
LSED	LSED is combined with several centrality methods separately to calculate localization-specific centrality scores for proteins based on the protein subcellular localization interaction networks.	Unweighted	PPIs; Subcellular localization information	[95]
UC	UC is a united complex centrality that integrates the protein complexes with the topological features of PPI networks.	Weighted	PPIs; Protein complex	[107]

Continued

Table 3. (continued)

Methods	Short Description	Network	Data	Reference
SON	SON integrates the information of subcellular localization, orthologous proteins and PPI networks by a linear combination strategy.	Weighted	PPIs; Subcellular localization information; Orthology information	[108]
CIC	CIC scores the edges taken place in different compartments.	Weighted	PPIs; Gene expression profiles; Subcellular localization information	[102]
PeC	PeC scores each edge in the PPI networks using NC and PCC.	Weighted	PPIs; Gene expression profiles	[103]
WDC	WDC assigns a suitable ratio for PCC and NC to evaluate the importance of edges in PPI networks.	Weighted	PPIs; Gene expression profiles	[104]
TS-PIN	TS-PIN keeps interactions that the two proteins appear in the same subcellular location and are activated at least at one time point.	Unweighted	PPIs; Gene expression profiles; Subcellular localization information	[123]
NF-APIN	NF-APIN uses time-dependent and time-independent models to process gene expression profiles and constructs an active PPI networks based on co-expressed genes.	Unweighted	PPIs; Gene expression profiles	[33]
Zhong *et al.*	Integrating multiple biological features and topological features into an SVM-RFE model.	Unweighted	PPIs; Gene expression profiles; Orthology information; Subcellular localization information	[126]
CENT-ING-GO	Integrating topological features and GO information into an ensemble model.	Unweighted	PPIs; Transcriptional regulatory networks; Metabolic networks; GO information	[34]
Deng *et al.*	Integrating topological features and sequences into an ensemble model.	Unweighted	Gene expression networks; Sequence data	[48]
Gustafson *et al.*	Integrating topological features and sequences into a naive bayes model.	Unweighted	PPIs; ORF length; Paralogs; Codon adaptation; Phyletic retention	[128]
Acencio *et al.*	Integrating cellular localization, topological features and BP information into a decision tree-based model.	Unweighted	PPIs; Cellular localization information; GO information	[129]
Zeng *et al.*	Integrating PPI topological features and gene expression profile patterns into a deep learning model.	Unweighted	PPIs; Gene expression profiles	[130]

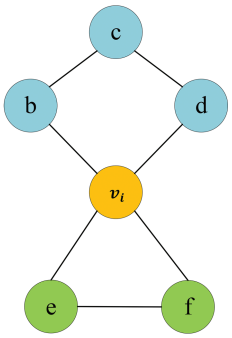

Figure 1. A schematic network demonstrating different topology-based methods. We use v_i (yellow node) as an example to demonstrate different topology-based methods. The DC of node v_i is 4 and its normalized DC value is 0.8, because it has four edges connecting with *b*, *d*, *e*, *f*. The LAC of node v_i is equal to 0.5, because $S_i = \{b, d, e, f\}$ and C_i only contains the edge between *e* and *f*. The NC of node v_i is equal to 2, because there is one triangle consisting of v_i, *e* and *f*. For the method of DMNC, *N* is equal to 2 and *E* is equal to 1, that is, *e*, *f* and the edge between them. The TP-based centrality of node v_i is equal to 1.33. The CC of node v_i is equal to 0.833, because the shortest paths of v_i and *b*, *c*, *d*, *e*, *f* are 1, 2, 1, 1 and 1, respectively. According to the formula, the IC of node v_i is 2.25, the BC of node v_i is 13, the SC of node v_i is 4.617 and the EC of node v_i is equal to 0.598.

proposed to reduce the false positives when predicting essential proteins with high connectivity from a different perspective. NC is a method that binds the importance of edges and the closeness of two nodes effectively. For a given node v_i, its NC value is defined as the following equation.

$$NC(i) = \sum_{j \in N_i} \frac{z_{i,j}}{\min\left(DC(i) - 1, DC(j) - 1\right)}. \quad (3)$$

For the edge $e_{i,j}$, $z_{i,j}$ denotes the number of triangles that contain the edge $e_{i,j}$ actually in the network. $DC(i)$ and $DC(j)$ correspond to the degree of nodes v_i and v_j, respectively, $\min\left(DC(i) - 1, DC(j) - 1\right)$ denotes the maximum number of triangles that the edge $e_{i,j}$ can participate in.

Density of maximum neighborhood component

Based on the conclusion that essential genes or proteins are inclined to appear in clusters, DMNC [68] identifies whether a gene or protein is essential or not by calculating the density among the neighbors of this protein. For a given node v_i, its DMNC value is defined as the following equation.

$$DMNC(i) = \frac{E(C(i))}{N(C(i))^{\varepsilon}}, \quad (4)$$

where $C(i)$ is a maximum neighborhood subnetwork composed of the neighbors of v_i and their edges, and $C(i)$ does not include the node v_i and the edges which is related to v_i. *N* is the number of nodes and *E* is the number of edges in $C(i)$. The range of

parameter ε is between 1 and 2. In general, ε is set to be 1.7, which can obtain the better prediction results.

Topology potential-based centrality

The concept of physics field was used to describe the non-contact interactions of material particles. Inspired by the development of field theory, TP [69] regards the PINs as a physical system and calculates the topology potential of each protein to analyze its essentiality in the network. For a given node v_i, its TP value is defined as the following equation.

$$TP(i) = \sum_{j=1}^{n} \left(m_j \times e^{-\left(\frac{\rho(i,j)}{\sigma}\right)^2} \right). \quad (5)$$

For a node v_j influenced by the node v_i, m_j is the quality of node $v_j (j = 1, \ldots, n)$ and the value is 1. $\rho(i,j)$ is the shortest path of v_i and v_j in the network. For the impact factor σ, since essential proteins has the characteristics of clustering, their influence range only includes direct neighbor nodes and indirect neighbor nodes [70, 71], the optimal impact factor of the PIN is set to 0.9428, thus ensuring that the influence range of a protein is not exceed 2.

TP-NC [69] is an improved method of TP, which utilizes NC as the intrinsic attribute of TP, that is, TP-NC calculates a protein's m_j with NC.

In addition to the above-mentioned methods, there are some other neighborhood-based methods that can also be used to predict essential proteins. For example, Kitsak *et al.* [72] propose the k-shell decomposition method to determine the location of nodes in the network. As the location of a node in network is also a crucial factor in predicting the essentiality of the node, the k-shell decomposition method can also be used to predict essential proteins. There are also some researchers who integrate or improve the above neighborhood-based methods. For example, Shang *et al.* [73] consider that LAC only takes into account the direct neighbors, they propose a new method named LAC2 that integrate the local connectivity values of nodes and their neighbors together.

Path-based methods

Unlike the methods based on neighborhood, path-based methods consider the global topological characteristics in the PINs. The typical path-based approaches are CC [74], information centrality (IC) [75], BC [76] and subgraph centrality (SC) [77].

Closeness centrality

In general, the central nodes of the network have less reliance on other nodes, whereas the nodes at the verge of the network must rely more on other nodes. CC [74] indicates how close the current node is to all other nodes and reflects the centrality of the node in the network. The greater the CC value of a node is, the shorter the average distance between this node and others is. It indicates that this node has low dependence on other nodes and it is more likely to be in the center of the network. CC of a node is inversely proportional to average shortest path from this node to all the other nodes. For a given node v_i, its CC value is defined as the following equation.

$$CC(i) = \frac{n-1}{\sum_{j \neq i} \rho(i,j)}, \quad (6)$$

where $\rho(i,j)$ denotes the shortest path of v_i and v_j in the network.

The disadvantage defined above is that it can only be used in the connected networks. Latora and Marchiori [78] improve the above method to enable it to be used in the unconnected networks.

$$EFF(i) = \sum_{j=1}^{n} \frac{1}{\rho(i,j)}. \quad (7)$$

If there is no path reachable between nodes v_i and v_j, then $\rho(i,j) = \infty$. CC uses the relative distance between all pairs of nodes to determine the centralities of nodes, which is widely used but has a high time complexity.

Information centrality

For a node v_i, IC [75] measures the harmonic average of all paths that end at v_i. IC value is defined as the following equation.

$$IC(i) = \left[\frac{1}{n} \sum_{j} \frac{1}{I_{ij}} \right]^{-1}. \quad (8)$$

Defining the matrix $C = (c_{ij}) = (D - A + J)^{-1}$, where all elements in the n-dimensional matrix J are 1. Matrix D is an n-dimensional diagonal matrix whose elements are the degree of all nodes in the network. Thus I_{ij} can be denoted as $I_{ij} = (C_{ii} + C_{jj} - 2C_{ij})^{-1}$. For the sake of computation, I_{ii} is considered as infinite, thus, $1/I_{ii}$=0.

IC reflects the information contained in all possible paths in a network. The information content is inversely proportion to the distance of a path in a network and the information in a combined path is equal to the sum of the information of the individual paths. Thus IC can simplify the computational process and can be easily extended to a weighted network as well as to an unconnected network.

Betweenness centrality

For the shortest paths of all node pairs, BC [76] considers that the more the shortest paths pass through a node are, the more important the node is. BC of a given node v_i is defined as

$$BC(i) = \sum_{k} \sum_{j} \frac{\rho(k,i,j)}{\rho(k,j)}, k \neq i \neq j \quad (9)$$

Where $\rho(k,i,j)$ corresponds to the number of shortest paths from node v_k to node v_j that pass through the node v_i, and $\rho(k,j)$corresponds to the number of the shortest paths from nodes v_k to v_j.

In some cases, the different nodes may have the same DC or CC values, then the importance of the nodes can be distinguished by the BC value [77]. In addition, BC can be used to characterize the 'modularity' of various biological and social networks.

Subgraph centrality

SC [77] of a node v_i calculates the number of closed loops in which v_i appears. SC is related to the length of the closed loop. The shorter the length of the closed loop, the more convenient

exchange of information on the closed loop. For a given node v_i, its SC value is defined as the following equation.

$$SC(i) = \sum_{l=0}^{\infty} \frac{\mu_l(i)}{l!} = \sum_{v=1}^{N} (\xi_v(i))^2 e^{\lambda_v}, \quad (10)$$

where $\mu_l(i)$ represents the number of closed loops of length l starting and ending at node v_i. λ_v is the eigenvalue of the adjacency matrix A and ξ_v is the eigenvector corresponding to λ_v. $\xi_v(i)$ is the ith component of ξ_v.

In addition, some other path-based methods, such as eccentricity centrality [79], bottleneck [80] can also be adopted to predict the essentiality of a node in network effectively. Eccentricity centrality of a node is defined as the maximum of distance from this node to all others in the network. Bottleneck considers that nodes with the highest BC values control the majority of information flow, which is regarded as the essential nodes.

Eigenvector-based methods

Neighborhood-based and path-based approaches identify the essentiality of a given node by the number of nodes that is connected to this node in networks. Eigenvector-based approaches not only consider the number of nodes that are connected to the node but also consider the influence of their qualities. Typical eigenvector-based measures are: EC [81], google's PageRank centrality (PR) [82], LeaderRank centrality (LR) [83] and hyperlink-induced topic search (HITs) [84].

Eigenvector centrality

A node's EC [81] represents the intensity of the influence of the node's neighbors on it. EC considers that the essentiality of a node depends on both the number of its neighbors and the importance of each neighbor. EC of a given node v_i is defined as the following equation.

$$EC(i) = \alpha_{max}(i), \quad (11)$$

where α_{max} is the principal eigenvector of the adjacency matrix A, and $\alpha_{max}(i)$ is the ith component of eigenvector α_{max}.

When there are some nodes with exceptional high degree (hubs) in the network, the localization phenomenon of EC will occur. That is, most of values are concentrated on the hubs, making values of other nodes very low. In order to avoid this phenomenon, Martin *et al.* [85] have proposed an improved EC method based on the nonbacktracking matrix, which can obtain values close to EC and eliminate the effect of localization.

Google's PageRank centrality

PR [82], a variant of EC, holds that the importance of a node depends on the quantity and quality of the other nodes pointing to it. For a given node v_i, the PR value can be calculated by the following equation.

$$PR_i(t) = (1-d)\sum_{j=1}^{n} \frac{a_{ji}}{k_j^{out}} PR_j(t-1) + \frac{d}{n}. \quad (12)$$

In each step, the PR value will be assigned to all nodes in the network with the probability of d and will be assigned to the nodes that are pointed by the node v_i with the probability of $1-d$. Let a_{ji}/k_j^{out} denotes the probability that a random walker goes from v_j to v_i in the next step, k_j^{out} is the out-degree of node v_j. The value of the parameter d depends on the specific situation and can be set between 0 and 1. The larger the value, the faster the convergence, while the lower the effectiveness of the algorithm. When $d = 1$, all nodes have the same PR value. When $d = 0$, The PR value of node v_i at time t is calculated by the following equation.

$$PR_i(t) = \sum_{j=1}^{n} \frac{a_{ji}}{k_j^{out}} PR_j(t-1). \quad (13)$$

However, the drawback of the above equation is that once a PR value reaches a node with a zero out-degree (called a dangling node), it will stay at that node forever and cannot be transmitted, thus it will continuously hoard the PR value [82, 86].

Researchers have proposed a series of improved algorithms based on PR. For example, in order to avoid the problem that dangling nodes hoard PR values, Kim and Lee [86] evenly distribute the PR values of dangling nodes to the n nodes in the network in each step. It is equally probable PR picks the next node from a node. However, Zhang *et al.* [87] consider that the greater the out-degree of the n nodes, the more likely they are picked, thus proposing the N-step PageRank algorithm.

LeaderRank centrality

LR [83] is proposed on the basis of PR. The parameter d in PR is replaced by the addition of a ground node that connects to all nodes through directional links, thus LR is parameter-free and adaptive. The existence of the ground node also ensures the strong connectivity of the network. For a given node v_i, its LR value at time t is defined as the following equation.

$$LR_i(t) = \sum_{j=1}^{n+1} \frac{a_{ji}}{k_j^{out}} LR_j(t-1). \quad (14)$$

In the iteration process, the dimension of adjacency matrix is $n+1$. At the steady state, the value of the ground node will be evenly assigned to all nodes. Subsequently, the final value of LR is defined as

$$LR_i = LR_i(t_c) + \frac{LR_g(t_c)}{n}, \quad (15)$$

where t_c is the convergence time and $LR_g(t_c)$ is the value of the ground node at the steady state.

Experiments have found that LR performs better than PR in some ways. (i) LR converges faster than PageRank [88], (ii) LR can better identify essential nodes in the network and (iii) LR is more robust than PR in resisting random interference.

Hyperlink-induced topic search

HITs centrality [84] gives two metric values for each node: authorities and hubs. Authorities measure the initial query of a node and hubs embody the importance of a node in the information dissemination. Defining a_i^t and h_i^t as the authority and hub of node v_i at time t, respectively.

$$a_i^t = \sum_{j=1}^{n} a_{ji} h_j^{t-1}, \quad h_i^t = \sum_{j=1}^{n} a_{ij} a_j^t. \quad (16)$$

The normalization process is required after the end of each step

$$a_j^{rt} = \frac{a_i^t}{\| a^t \|}, \quad h_i^{rt} = \frac{h_i^t}{\| h^t \|}. \tag{17}$$

HITs are the first method to sort nodes in the network simultaneously with different metrics, it can be used to determine multiple interrelated attributes of a node in network and handle more complex ranking problem.

Besides the above topology-based methods that calculate the centralities of a node from the global, local or eigenvector perspective, there are some methods that combine different types of topological centralities together. Rio *et al.* [89] have compare the reliability of essential genes identified by 16 different topology-based methods in 18 different reconstructed metabolic networks for yeast, finding that a single topology-based method does not significantly identify the essential genes, but methods that combine at least two topological centralities can improve prediction accuracy effectively.

All the above studies have shown that the topological centralities are tightly bound to the essentiality of genes/proteins. However, it is still a big challenge to further study the intrinsic relationship of a protein's topological characteristic and its essentiality and improve the prediction accuracy of the network-based methods as the high positives and false negatives in current PINs. It has been shown that not all essential genes/proteins are highly connected and some of essential proteins are low connectivity. Hence, we need to analyze the essential proteins' topological characters from different types of network centralities [90–93].

Integrating PINs and biological information to predict essential genes or proteins

In order to overcome the above-mentioned constraints and improve the accuracy of prediction, some researchers predict essential genes/proteins by integrating PINs and biological information. These methods not only utilize multiple networks such as PINs, gene regulatory networks and metabolic networks, but also combine a large number of biological data, such as microarray data, RNA-seq data, subcellular localization information, protein complex information, orthology information, protein domains and other known functional information. The predicted proteins are subsequently evaluated by comparison to the known essential proteins. The schematic diagram of integrating PINs and biological information to predict essential genes or proteins is shown in Figure 2.

The existing prediction methods can be divided into the following three perspectives: filtering ineffective interactions, weighting edges and some other network-based methods. Table 3 lists the summary of methods that integrate PINs and biological information.

To reduce the effects of noise such as false positives and negatives, some methods filter ineffective interactions in PINs. Li *et al.* [94] utilize the gene expression and subcellular localization information to refine the neighbors of proteins and then harness a random walk algorithm to integrate proteins' topological centralities and their orthologous information. Based on the idea that there is a greater possibility of physical interaction between two proteins if they present in the same subcellular location and are active at least at one time point in the cell cycle, Peng *et al.* [95] reconstruct PINs based on subcellular location information and the centrality–lethality theory is rechecked. Additionally, Li *et al.* [96] partition the original network by analyzing the distribution characteristics of proteins in different subcellular structures, so as to obtain subnetworks with different priorities.

Some methods improve the accuracy of prediction by weighting edges. Li *et al.* [97] calculate the confidence score of every edge in the PINs by combining the logistic regression model and functional similarity methods using GO semantic similarity and PPI data, then a weighted network has been built based on the scores to identify essential proteins. By analyzing the number of orthologs of yeast proteins in 99 reference species such as Homo sapiens (H.sapiens) and Escherichia coli (E.coli), Peng *et al.* [98] find that the more frequently a protein appears in reference species, the more likely that the protein is the essential protein and propose the iteration method named ION to weight edges by combining the homologous information and NC. Fan *et al.* [99] find that the combination of multiple data sources can improve the predictive performance; in particular, subcellular localization information can facilitate the identification of essential proteins. Consequently, a method named SCP has been proposed, which combines both subcellular localization information and Pearson correlation coefficient (PCC) in a modified PR algorithm. On the basis of the integration of modularity and conservatism, Zhao *et al.* [100] build three individual weighted networks, which utilize the topological features of PINs, protein domain, gene expression and orthologous information of proteins. Peng *et al.* [101] combine protein domains and topological properties based on the discovery that proteins are inclined to be indispensable if they contain more domain types, while other proteins do not. They weight edges in PINs by NC and the essential probability of the two proteins that are calculated according to the quantity and frequency of their domain type, then the importance of a protein can be obtained by computing the sum of values of its adjacent edges. Lei *et al.* [32] have constructed a weighted PINs and proposed an algorithm RSG based on the GO terms information, subcellular compartments and RNA-seq data. According to the assumption that edges in the PIN are of different importance in different subcellular locations, Peng *et al.* [102] propose a method named CIC to score the edges taken place in different compartments. Moreover, Li *et al.* [103] consider that essential genes or proteins are prone to form densely connected clusters, and essential genes or proteins in the same cluster are more easily co-expressed. Thus, they propose a method named PeC to score each edge in the PIN using NC and PCC. Similarly, Tang *et al.* [104] assign a suitable ratio for PCC and NC to evaluate the importance of edges in the PIN.

Additionally, Luo *et al.* [105] propose an algorithm named LIDC by considering LID centrality and in-degree information of protein complexes. Zhang *et al.* [106] come up with an ensemble framework integrating PINs and gene expression data to improve the accuracy of typical topological measures. By comparing the known essential proteins and the known protein complexes, Li *et al.* [107] find that there is a close correlation between proteins in complexes and the importance of proteins. That is, proteins in complexes have a greater probability of being indispensable than proteins not in any complex, and proteins found in multiple complexes are more indispensable than those found in only one complex. Therefore, the algorithm UC combines protein complexes and topological centralities together to predict essential proteins. Since essential proteins are more inclined to be found in certain subcellular locations and their evolution is more conservative, Li *et al.* [108] utilize orthology and subcellular localization data to predict essential proteins. Additionally, Li *et al.* [109] propose a scheme based on priori knowledge and integrate gene expression profiles and network topology in this new scheme to identify essential proteins.

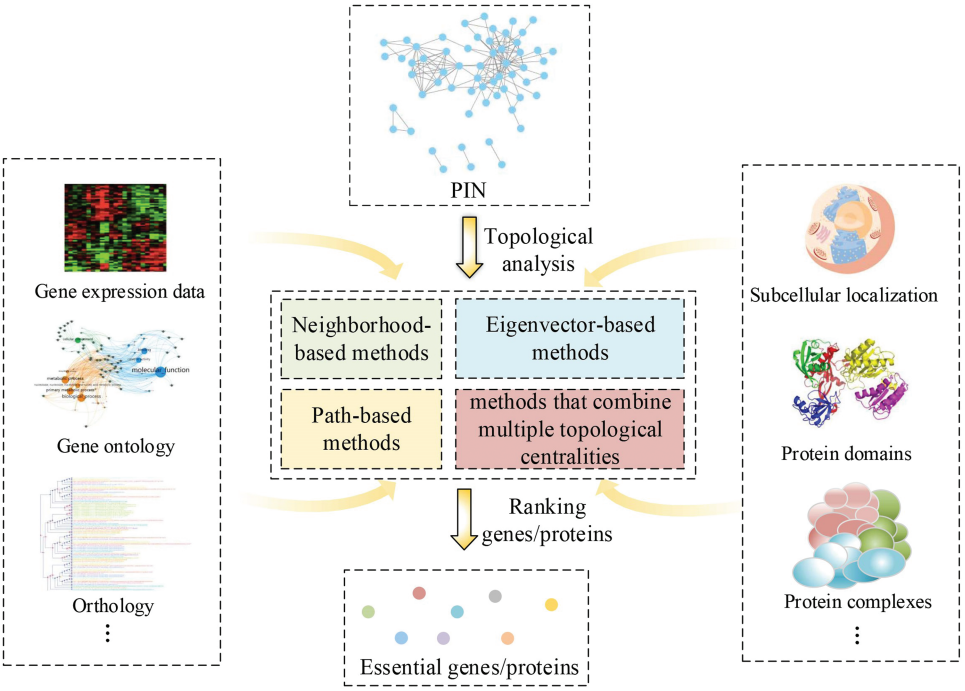

Figure 2. Schematic diagram of integrating PINs and biological information to predict essential genes or proteins.

Dynamic network-based methods to predict essential genes or proteins

PINs change with time and conditions to ensure normal life activities. The dynamic of cells in time and space plays a vital role in the replication and viability of organisms [111–114]. Traditional methods construct static networks to analyze complex networks under a single condition, but static networks cannot reveal the dynamic of PINs. With the occurrence and development of multi-omics data, it has become possible to construct more comprehensive dynamic networks to predict essential genes or proteins (Figure 3).

Recently, many studies identify the activities of genes or proteins by analyzing the gene expression level at a certain time and constructing a time series dynamic PIN to characterize the dynamic of complex biological systems [115–117]. According to DNA microarray time series, Lichtenberg *et al.* [118] construct a time series dynamic network based on both periodically expressed genes with peak expression at different time points and the protein interactions provided by the STRING database. However, the time series dynamic networks constructed by the periodically expressed genes with peak expression are small in size, resulting in the loss of valuable biological information and the inability to fully describe the complex process of PINs with time. Hegde *et al.* [119] mention that the noise in gene expression data is inversely proportional to the average gene expression level. Only the expression levels of genes not lower than the average values can be considered as valid expression and can greatly eliminate the effects of noise.

However, setting a fixed threshold cannot be convincing considering that there are also low-expressed proteins that are significant in the cell cycle. In order to properly determine the activities of proteins, a 3-sigma principle has been proposed to calculate the dynamic threshold of each gene according to the distribution characteristics of gene expression levels [120]. Meng *et al.* [121] use the 3-sigma principle to determine the activated time of proteins and consider the dynamic spatial information provided by subcellular localization of proteins to construct a spatial and temporal active PIN. Shen *et al.* [122] think that the 3-sigma principle will lead a lot of high-expression proteins to be filtered out, resulting in the incomplete analysis of dynamic interactions. In order to solve this problem, they screen active proteins based on the deviation degree of gene expression curves and build a time-evolving PIN to simulate the dynamic process of protein interactions.

Dynamic networks have been successfully applied in the identification of essential genes or proteins. Li *et al.* [123] have purified the PINs by considering that a protein interaction is valid when the two proteins appear in the same subcellular location and are activated at least at one time point. Subsequently, the prediction accuracy can be improved by using the refined PIN. Xiao *et al.* [33] divide the time series gene expression data into two categories: time-dependent and time-independent data. If the expression level of a gene at each time point is time-independent and the average gene expression value is small, the gene will be judged as noise and filtered out. Then they use the 3-sigma principle to determine the time of protein activity and construct an active PIN to predict essential proteins. The summary of dynamic network-based methods is listed in Table 3.

Machine learning-based methods to predict essential genes or proteins

The identification of essential genes or proteins can be regarded as a binary classification problem, its purpose is to classify a list of genes or proteins into two groups by suitable features, which are closely related to the essentiality of genes or proteins. There are three criteria for feature extraction. First, the features should be easy to obtain. Second, the features should be powerful to

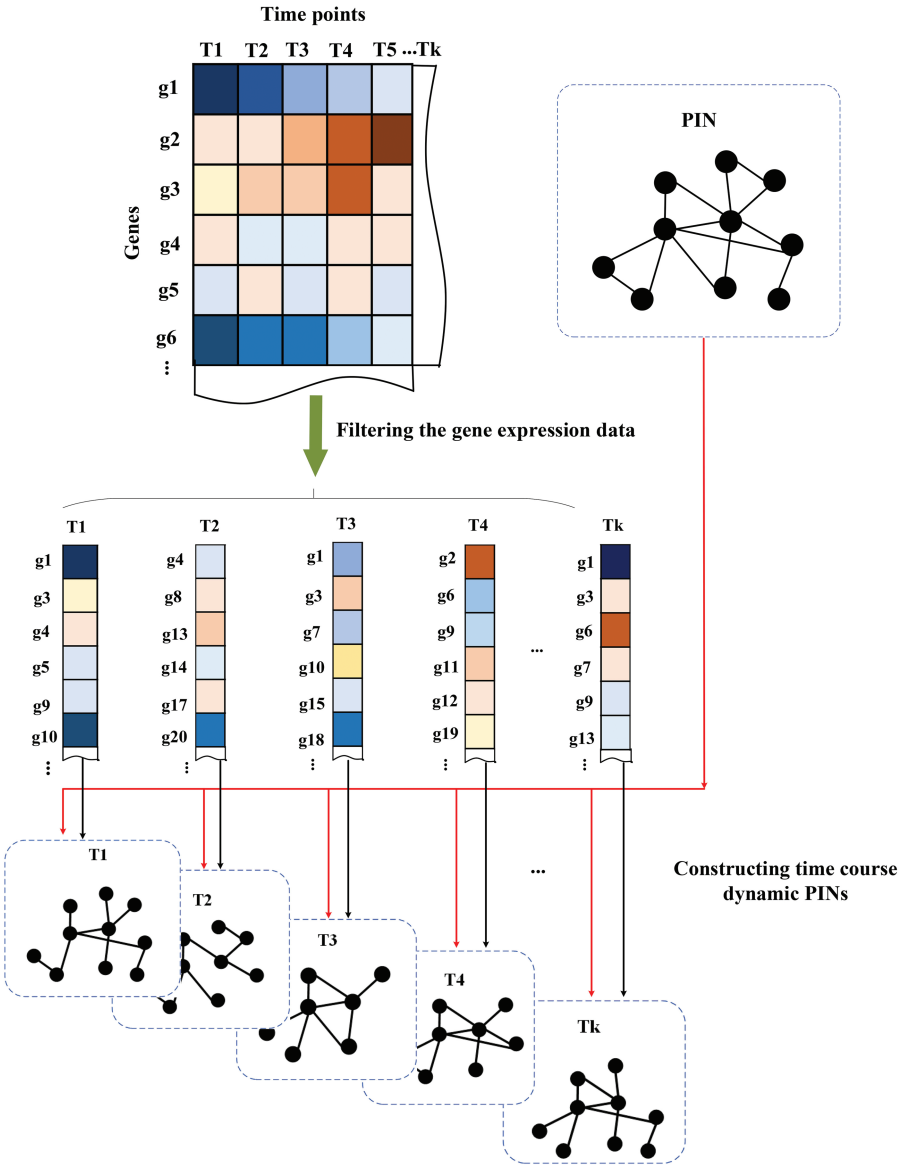

Figure 3. Schematic diagram of the time series dynamic PINs construction. First, according to the gene expression information, it is determined whether a protein is expressed at the time points and under-expressed genes are filtered out. Second, constructing time series dynamic PINs combining static PIN and microarray time series [110].

identify essential genes or proteins. Third, the features should minimize biological redundancy [124].

Nowadays, the use of machine learning methods to predict essential proteins has become a hot topic [125]. The commonly used models include support vector machine (SVM), decision tree, neural network, naïve Bayes and so on. The summary of machine learning-based methods is listed in Table 3.

In order to make the extracted features meet the aforementioned criteria, some researchers have analyzed the relationship between essentiality and features and the correlation between the features. On the basis of the collection of 26 features, Zhong *et al.* [126] select suitable features by adopting support vector machines-recursive feature elimination (SVM-RFE) with 10-fold cross validation to rank the feature list and select the suitable subset of features. Subsequently, PCC method has been used to evaluate the correlations between features in the subset. After removing certain features that have the higher PCC values, the resulting features not only have powerful prediction ability for protein essentiality but also share minimal biological meaning with each other. Since the features from experimental omics data are often not available, Guo *et al.* [127] only use the inherent characteristics derived from sequences, from which nucleotide composition and association information can be denoted using a λ-interval Z-curve. Then the SVM classification model is constructed to identify human essential proteins. Kim [34] uses GO information to measure the similarity between proteins and purifies the PINs by removing a number of 'biologically insignificant' edges. Subsequently, eight centrality measures are calculated on the refined PINs and 23 features are extracted. The classifiers are trained by using seven decision-tree based models, SVM and a neural network model, and then a 'vote' method is used to classify the proteins.

Deng *et al.* [48] adopt four separate classifiers (naive Bayes classifier, logistical regression model, C4.5 decision tree and CN2

Table 4. Tools for essential gene/protein prediction based on PINs

Tools	Centralities	Platform	URL	Reference
Cyto-Hubba	Bottleneck; Maximum neighborhood component; DMNC; Double Screening scheme of MNC and DMNC; Edge percolated component	Cytoscape	http://hub.iis.sinica.edu.tw/cytohubba/	[131]
CytoNCA	DC; CC; EC; IC; LAC; NC; SC; BC	Cytoscape	http://apps.cytoscape.org/apps/cytonca	[132]
CentiLib	DC; BC; Centroid centrality; CC; Current-flow CC; Current-flow BC; Eccentricity centrality; EC; HITS; Hubbel index; Katz status index; PageRank; Radiality centrality; Stress centrality	Vanted and Cytoscape	https://immersive-analytics.infotech.monash.edu/centilib/	[133]
CentiScaPe	BC; Bridging centrality; Centroid centrality; CC; Eccentricity centrality; EC; Radiality centrality; Stress centrality	Cytoscape	http://www.cbmc.it/~scardonig/centiscape/centiscape.php	[134]
NetworkAnalyzer	DC; BC; CC; Clustering coefficient; Eccentricity centrality; Neighborhood connectivity; Radiality centrality; Shared neighbors; Shortest paths; Stress centrality; Topological coefficients	Cytoscape	http://med.bioinf.mpi-inf.mpg.de/networkanalyzer/	[135]
SBEToolbox	BC; Bridging centrality; CC; Clustering coefficient; DC; Eccentricity centrality; Knotty centrality	MatLab toolbox	https://github.com/biocoder/SBEToolbox/releases	[136]
FUGA	DC; EC; Clustering coefficient; Eccentricity centrality; Local efficiency; Node BC	MatLab toolbox	https://code.google.com/archive/p/fuga/	[137]
CentiBiN	DC; Eccentricity centrality; Bargaining centrality; Centroid centrality; CC; Closeness vitality; Current-flow BC; Current-Flow CC; EC; HITS; Hubbell index; Katz status index; PageRank; Radiality centrality; Shortest-Paths BC; Stress centrality	Java standalone platform	http://centibin.ipk-gatersleben.de/	[138]
DyNetViewer	TC-PIN, DPIN, NF-APIN; ST-APIN; BC; CC; DC; EC; LAC; NC; SC; IC; Stress centrality; Radiality centrality; Eccentricity centrality; Centroid centrality	Cytoscape	http://apps.cytoscape.org/apps/dynetviewer	[139]

rule) to evaluate the predictive ability of different features for protein essentiality. They sort all features based on the coverage length of log-odds ratio. The longer the total coverage length is, the more relevant the corresponding feature is to the essentiality of proteins. Those with positive effect and a monotonic relationship with essentiality are the candidate features. Then, they minimize biological meaning by removing candidate features that have the higher PCC values. Gustafson *et al.* [128] integrate topological features and apply a naive Bayes classifier to measure the predictive performance of both the individual feature and the integrated features.

Acencio *et al.* [129] extract 12 different network topological characteristics, and biological characteristics such as BP terms in GO are also adopted. Based on these characteristics, a decision tree-based meta-classifier is trained and tested to identify essential proteins.

Recently, deep learning methods have been successfully applied in the field of essential gene/protein identification. Zeng *et al.* [130] propose a deep learning framework that makes use of the node2vec technique to learn topological features from PPI network and utilize convolutional neural networks to extract the patterns of gene expression profiles.

Tools and applications

At present, a variety of tools have been developed to predict essential genes or proteins based on PINs. Here, we introduce some useful tools that integrate multiple network-based methods for essential gene prediction: Cyto-Hubba [131], CytoNCA [132], CentiLib [133], CentiScaPe [134], NetworkAnalyzer [135], SBEToolbox [136], FUGA [137], CentiBiN [138] and DyNetViewer [139]. The details of these tools are listed in Table 4.

Cyto-Hubba provides the analysis of node essentiality in biological networks and subnetworks composed of essential nodes. CytoNCA provides computation, evaluation and visualization analysis for several centrality indexes of weighted and unweighted network. CentiLib calculates weighted and unweighted centralities in biological networks. CentiScaPe integrates the computation of centralities for undirected, directed and weighted networks. NetworkAnalyzer calculates specific topological parameters of molecular interaction networks and visualizing the results. SBEToolbox calculates centralities and topological statistics for biological networks and it only supports undirected networks. FUGA is a toolbox for inference and analysis of biological and cellular networks. CentiBiN provides 17 topology-based methods for directed or undirected networks and the visualization of centralities. DyNetViewer provides four methods for the construction of dynamic networks and topology-based methods to analyze the essentiality of nodes in dynamic networks.

Challenges and future work

Essential genes are necessary for the survival of organisms and their products, essential proteins, play an important role in the growth and development of organisms because of their unique biological functions. Computational methods for identifying essential genes and proteins can reduce the workload and provide new candidate genes and proteins for biologists. In this paper, we review the advanced methods for predicting essential

genes/proteins based on network. For making a more informed judgment about the utility of these methods, we compare them on six yeast PINs that include two PINs with different sizes and four PINs with different reliability (Supplementary Materials). The two different size PINs are obtained from DIP and BIOGRID, respectively and other four PINs are collected from Mering *et al.* [140], named Y2K, Y11K, Y45K and Y78K according to the credibility, respectively. The results indicate that the network-based methods do have the ability to efficiently predict essential genes or proteins. Meanwhile, these methods can obtain better prediction results on the highly reliable networks such as Y2K than the predictions on high-noise networks. With the increase of the scale and noise of networks, the predictive performance of some methods will reduce, while some methods integrating biological information have the anti-noise ability and still maintain the predictive ability. Therefore, integrating effective biological information and constructing highly reliable networks provides an effective way to improve the accuracy of essential gene/protein predictions.

In what follows, we present the challenges and future work.

The PPI data measured by experimental techniques is incomplete and has noise, which will reduce the accuracy of network-based methods for predicting essential genes or proteins. Therefore, the big challenges of further researches are how to present effective pre-processing methods to process PPI data and develop suitable pre-processing techniques for each organism. At the same time, in addition to PINs, other networks such as metabolic networks [141], signal transduction networks [142] and gene regulatory networks [143, 144] have also been proved to be related to essential genes [145, 146]. How to predict essential genes in combination with multiple networks is the focus of future work.

PINs are dynamic, which is embodied in two aspects: temporal dynamics and spatial distribution characteristics. Current researches mainly analyze the dynamic changes of genes or proteins in time series based on the microarray data. However, with the development of next-generation sequencing technologies such as RNA-seq, a more comprehensive gene expression map is provided. It is necessary to collect and sort RNA-seq time series data to design a new dynamic network model. Moreover, single-cell sequencing data can reflect the intracellular network, which can promote the prediction of essential genes or proteins. The effective combination of single-cell sequencing data to build more reliable networks is the focus of future research. Meanwhile, single-cell proteomics [147] may become a reality and be used to build more accurate PINs. Furthermore, PINs will change with different tissues or subcellular location, which indicates that spatial information is also important to construct dynamic networks. Consequently, the future work is how to combine spatial information and design reasonable and effective methods to determine the active PINs.

With the development of high-throughput technologies, a great deal of omics data such as genomics, proteomics and transcriptomics has been accumulated, while the relationship between some biological information and essential genes/proteins has not been thoroughly explored. The effective integration of multi-omics data to predict the essentiality of genes/proteins needs to be improved.

In most of the studies, model organisms, especially yeast, are used as the main research object [148, 149], while other species are poorly analyzed, such as various microbial flora. Meanwhile, identification of essential genes in human cancer cell lines is the valuable research direction. Furthermore, some studies have shown that essential genes or proteins tend to be evolutionarily conserved [150–152], which indicates that there is a close relationship between essential genes or proteins in different species. Predicting essential genes or proteins across species may be the focus of future work.

Key Points

- Available databases related to essential genes and proteins are introduced, and the popular tools are described.
- This article summarized the commonly used methods including topology-based methods, methods of integrating PINs and biological information, dynamic network-based methods and machine learning-based methods in predicting essential genes or proteins.
- Focusing on the difference between the various methods, we discussed the use of parameters of various methods and different applicable situations.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported in part by the National Natural Science Foundation of China (grant nos. 61832019, 61622213 and 61728211), the 111 Project (grant no. B18059) and the Hunan Provincial Science and Technology Program (2018WK4001).

References

1. Peng C, Lin Y, Luo H, *et al.* A comprehensive overview of online resources to identify and predict bacterial essential genes. *Front Microbiol* 2017;**8**:2331.
2. Glass JI, Hutchison CA, Smith HO, *et al.* A systems biology tour de force for a near-minimal bacterium. *Mol Syst Biol* 2009;**5**:330–2.
3. Koonin EV. How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* 2000;**1**:99–116.
4. Lamichhane G, Zignol M, Blades NJ, *et al.* A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* 2003;**100**:7213–8.
5. Park D, Park J, Park SG, *et al.* Analysis of human disease genes in the context of gene essentiality. *Genomics* 2008;**92**:414–8.
6. Furney SJ, Albà MM, López-Bigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics* 2006;**7**:165.
7. Wang J, Chen G, Li M, *et al.* Integration of breast cancer gene signatures based on graph centrality. *BMC Syst Biol* 2011;**5**:S10.
8. Giaever G, Chu AM, Ni L, *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 2002;**418**:387–91.
9. Chen L, Ge X, Xu P. Identifying essential Streptococcus sanguinis genes using genome-wide deletion mutation. *Methods Mol Biol* 2015;**1279**:15–23.

10. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunol Cell Biol* 2005;**83**:217–23.
11. Kamath RS, Fraser AG, Dong Y, *et al*. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 2003;**421**:231–7.
12. Ji Y, Zhang B, Van SF, *et al*. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 2001;**293**:2266–9.
13. Gallagher LA, Ramage E, Jacobs MA, *et al*. A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proc Natl Acad Sci U S A* 2007;**104**:1009–14.
14. Langridge GC, Phan M-D, Turner DJ, *et al*. Simultaneous assay of every Salmonella typhi gene using one million transposon mutants. *Genome Res* 2009;**19**:2308–16.
15. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. *Nature* 2002;**420**:218–23.
16. Alon U. Biological networks: the tinkerer as an engineer. *Science* 2003;**301**:1866–7.
17. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001;**98**:4569–74.
18. Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;**415**:180–3.
19. Ehrenberger T, Cantley LC, Yaffe MB. Computational prediction of protein–protein interactions. *Methods Mol Biol* 2015;**38**:1–17.
20. Rao VS, Srinivas K, Sujini G, *et al*. Protein–protein interaction detection: methods and analysis. *Int J Proteomics* 2014;**2014**:35–46.
21. Li M, Gao H, Wang J, *et al*. Control principles for complex biological networks. *Briefings in bioinformatics* 2018. 10.1093/bib/bby088.
22. Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. *Nature* 2000;**406**:378–91.
23. Jeong H, Mason SP, Barabási A-L, *et al*. Lethality and centrality in protein networks. *Nature* 2001;**411**:41–2.
24. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
25. Wagner A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 2001;**18**:1283–92.
26. Yu H, Greenbaum D, Lu HX, *et al*. Genomic analysis of essentiality within protein networks. *Trends Genet* 2004;**20**:227–31.
27. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 2004;**22**:803–6.
28. Wuchty S. Interaction and domain networks of yeast. *Proteomics* 2002;**2**:1715–23.
29. 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.
30. 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.
31. Lü L, Chen D, Ren X-L, *et al*. Vital nodes identification in complex networks. *Phys Rep* 2016;**650**:1–63.
32. Lei X, Zhao J, Fujita H, *et al*. Predicting essential proteins based on RNA-Seq, subcellular localization and GO annotation datasets. *Knowl-Based Syst* 2018;**151**:136–48.
33. Xiao Q, Wang J, Peng X, *et al*. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. *BMC Genomics* 2015;**16**:S1.
34. Kim W. Prediction of essential proteins using topological properties in GO-pruned PPI network based on machine learning methods. *Tsinghua Sci Technol* 2012;**17**:645–58.
35. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* 2008;**37**:D455–8.
36. Mewes H-W, Frishman D, Gruber C, *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2000;**28**:37–40.
37. Cherry JM, Adler C, Ball C, *et al*. SGD: Saccharomyces genome database. *Nucleic Acids Res* 1998;**26**:73–9.
38. Chen W-H, Lu G, Chen X, *et al*. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* 2016;**45**:D940–4.
39. Wattam AR, Davis JJ, Assaf R, *et al*. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* 2016;**45**:D535–42.
40. Lin Y, Zhang RR. Putative essential and core-essential genes in Mycoplasma genomes. *Sci Rep* 2011;**1**:53.
41. Xenarios I, Salwinski L, Duan XJ, *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002;**30**:303–5.
42. Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2010;**39**:D561–8.
43. Chatr-Aryamontri A, Oughtred R, Boucher L, *et al*. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 2017;**45**:D369–79.
44. Keshava Prasad T, Goel R, Kandasamy K, *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2008;**37**:D767–72.
45. Mewes H-W, Dietmann S, Frishman D, *et al*. MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 2008;**36**:D196–201.
46. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2011;**40**:D857–61.
47. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2011;**40**:D841–6.
48. Deng J, Deng L, Su S, *et al*. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res* 2010;**39**:795–807.
49. Finn RD, Bateman A, Clements J, *et al*. Pfam: the protein families database. *Nucleic Acids Res* 2013;**42**:D222–30.
50. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
51. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178–89.
52. Altenhoff AM, Glover NM, Train C-M, *et al*. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* 2017;**46**:D477–85.
53. Park D, Singh R, Baym M, *et al*. IsoBase: a database of functionally related proteins across PPI networks. *Nucleic Acids Res* 2010;**39**:D295–300.

54. Östlund G, Schmitt T, Forslund K, *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 2009;**38**:D196–203.
55. Zdobnov EM, Tegenfeldt F, Kuznetsov D, *et al*. OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* 2016;**45**:D744–9.
56. Huerta-Cepas J, Szklarczyk D, Forslund K, *et al*. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2015;**44**:D286–93.
57. Meinken J, Asch DK, Neizer-Ashun KA, *et al*. FunSecKB2: a fungal protein subcellular location knowledgebase. *Comput Mol Biol* 2014;**4**:1–17.
58. Lum G, Meinken J, Orr J, *et al*. PlantSecKB: the plant secretome and subcellular proteome knowledgebase. *Comput Mol Biol* 2014;**4**:1–17.
59. Meinken J, Walker G, Cooper CR, *et al*. MetazSecKB: the human and animal secretome and subcellular proteome knowledgebase. *Database* 2015;**2015**:1–14.
60. Rastogi S, Rost B. LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. *Nucleic Acids Res* 2010;**39**:D230–4.
61. Binder JX, Pletscher-Frankild S, Tsafou K, *et al*. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* 2014;**2014**:bau012.
62. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. *Nature Genet* 2000;**25**:25–9.
63. Consortium GO. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res* 2016; **45**:D331–8.
64. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 2003;**100**:12123–8.
65. Ren J, Wang J, Li M, *et al*. Discovering essential proteins based on PPI network and protein complex. *Int J Data Min Bioinform* 2015;**12**:24–43.
66. Pu S, Wong J, Turner B, *et al*. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 2008;**37**:825–31.
67. Ruepp A, Waegele B, Lechner M, *et al*. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res* 2009;**38**:D497–501.
68. Lin C-Y, Chin C-H, Wu H-H, *et al*. Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology. *Nucleic Acids Res* 2008;**36**: W438–43.
69. 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.
70. Hart GT, Lee I, Marcotte EM. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* 2007; **8**:236.
71. Li M, J-e C, J-x W, *et al*. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics* 2008;**9**:398.
72. Kitsak M, Gallos LK, Havlin S, *et al*. Identification of influential spreaders in complex networks. *Nat Phys* 2010;**6**: 888–93.
73. Shang X, Wang Y, Chen B. Identifying essential proteins based on dynamic protein–protein interaction networks and RNA-seq datasets. *Sci China Inform Sci* 2016;**59**: 070106.1–070106.11.
74. Wuchty S, Stadler PF. Centers of complex networks. *J Theor Biol* 2003;**223**:45–53.
75. Stephenson K, Zelen M. Rethinking centrality: methods and examples. *Soc Networks* 1989;**11**:1–37.
76. Joy MP, Brock A, Ingber DE, *et al*. High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol* 2005;**2005**:96–103.
77. Estrada E, Rodriguez-Velazquez JA. Subgraph centrality in complex networks. *Phys Rev E* 2005;**71**:056103.1–056103.9.
78. Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett* 2001;**87**:3–6.
79. Hage P, Harary F. Eccentricity and centrality in networks. *Soc Networks* 1995;**17**:57–63.
80. Yu H, Kim PM, Sprecher E, *et al*. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007;**3**:e59.
81. Bonacich P. Power and centrality: a family of measures. *Am J Sociol* 1987;**92**:1170–82.
82. Brin S, Page L. Reprint of: the anatomy of a large-scale hypertextual web search engine. *Comput Netw* 2012;**56**:3825–33.
83. Lü L, Zhang Y-C, Yeung CH, *et al*. Leaders in social networks, the delicious case. *PloS One* 2011;**6**:e21202.
84. Kleinberg JM. Authoritative sources in a hyperlinked environment. *J ACM* 1999;**46**:604–32.
85. Martin T, Zhang X, Newman M. Localization and centrality in networks. *Phys Rev E* 2014;**90**:052808.1–052808.7.
86. Kim SJ, Lee SH. An improved computation of the pagerank algorithm. In: *European Conference on Information Retrieval, 2002*, Vol. 2291, pp. 73–85. Springer, Berlin, Heidelberg.
87. Zhang L, Qin T, Liu T-Y, *et al*. N-step PageRank for web search. In: *European Conference on Information Retrieval, 2007*, Vol. 4425, pp. 653–60. Springer, Berlin, Heidelberg.
88. Li Q, Zhou T, Lü L, *et al*. Identifying influential spreaders by weighted LeaderRank. *Physica A Stat Mech Appl* 2014;**404**:47–55.
89. Del Rio G, Koschützki D, Coello G. How to identify essential genes from molecular networks? *BMC Syst Biol* 2009;**3**:102.
90. Chua HN, Tew KL, Li X-L, *et al*. A unified scoring scheme for detecting essential proteins in protein interaction networks. In: *2008 20th IEEE International Conference on Tools with Artificial Intelligence, 2008*, Vol. 2, pp. 66–73. IEEE, Piscataway, NJ.
91. Li M, Wang J, Wang H, *et al*. Essential proteins discovery from weighted protein interaction networks. In: *International Symposium on Bioinformatics Research and Applications, 2010*, Vol. 6053, pp. 89–100. Springer, Berlin, Heidelberg.
92. He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genetics* 2006;**2**:e88.
93. Zotenko E, Mestre J, O'leary DP, *et al*. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* 2008;**4**:e1000140.
94. 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.
95. Peng X, Wang J, Wang J, *et al*. Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks. *PloS One* 2015;**10**:e0130743.
96. Li M, Li W, Wu F-X, *et al*. Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. *J Theor Biol* 2018;**447**:65–73.

97. Li M, Wang J-X, Wang H, *et al*. Identification of essential proteins from weighted protein–protein interaction networks. *J Bioinform Comput Biol* 2013;**11**:1341002.1–1341002.19.
98. 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**:87.
99. Fan Y, Hu X, Tang X, *et al*. A novel algorithm for identifying essential proteins by integrating subcellular localization. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2016*, pp. 107–10. IEEE, Piscataway, NJ.
100. Zhao B, Wang J, Li X, *et al*. Essential protein discovery based on a combination of modularity and conservatism. *Methods* 2016;**110**:54–63.
101. Peng W, Wang J, Cheng Y, *et al*. UDoNC: an algorithm for identifying essential proteins based on protein domains and protein–protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2015;**12**:276–88.
102. Peng X, Wang J, Zhong J, *et al*. An efficient method to identify essential proteins for different species by integrating protein subcellular localization information. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2015*, pp. 277–80. IEEE, Piscataway, NJ.
103. Li M, Zhang H, J-x W, Pan Y. A new essential protein discovery method based on the integration of protein–protein interaction and gene expression data. *BMC Syst Biol* 2012;**6**:15.
104. Tang X, Wang J, Zhong J, *et al*. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**:407–18.
105. Luo J, Qi Y. Identification of essential proteins based on a new combination of local interaction density and protein complexes. *PloS One* 2015;**10**:e0131418.
106. Zhang X, Xiao W, Acencio ML, *et al*. An ensemble framework for identifying essential proteins. *BMC Bioinformatics* 2016;**17**:322.
107. Li M, Lu Y, Niu Z, *et al*. United complex centrality for identification of essential proteins from PPI networks. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:370–80.
108. Li G, Li M, Wang J, *et al*. Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinformatics* 2016;**17**:279.
109. 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.
110. Tang X, Wang J, Liu B, *et al*. A comparison of the functional modules identified from time course and static PPI network data. *BMC Bioinformatics* 2011;**12**:339.
111. Cohen AA, Geva-Zatorsky N, Eden E, *et al*. Dynamic proteomics of individual cancer cells in response to a drug. *Science* 2008;**322**:1511–6.
112. Przytycka TM, Singh M, Slonim DK. Toward the dynamic interactome: it's about time. *Brief Bioinform* 2010;**11**:15–29.
113. Hegele A, Kamburov A, Grossmann A, *et al*. Dynamic protein–protein interaction wiring of the human spliceosome. *Mol Cell* 2012;**45**:567–80.
114. Ren Z-M, Zeng A, Zhang Y-C. Structure-oriented prediction in complex networks. *Phys Rep* 2018;**750**:1–51.
115. Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 2001;**29**:3513–9.
116. Ge H, Liu Z, Church GM, *et al*. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* 2001;**29**:482–6.
117. Bhardwaj N, Lu H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics* 2005;**21**:2730–8.
118. De Lichtenberg U, Jensen LJ, Brunak S, *et al*. Dynamic complex formation during the yeast cell cycle. *Science* 2005;**307**:724–7.
119. Hegde SR, Manimaran P, Mande SC. Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol* 2008;**4**:e1000237.
120. Wang J, Peng X, Li M, *et al*. Active protein interaction network and its application on protein complex detection. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2011*, pp. 37–42. IEEE, Piscataway, NJ.
121. Meng X, Li M, Wang J, *et al*. Construction of the spatial and temporal active protein interaction network for identifying protein complexes. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2016*, pp. 631–6. IEEE, Piscataway, NJ.
122. Shen X, Yi L, Jiang X, *et al*. Mining temporal protein complex based on the dynamic pin weighted with connected affinity and gene co-expression. *PloS One* 2016;**11**:e0153967.
123. Li M, Ni P, Chen X, *et al*. Construction of refined protein interaction network for predicting essential proteins. *IEEE/ACM Trans Comput Biol Bioinform* 2017. 10.1109/TCBB.2017.2665482.
124. Wang J, Peng W, Wu F-X. Computational approaches to predicting essential proteins: a survey. *Proteomics Clin Appl* 2013;**7**:181–92.
125. Zhang X, Acencio ML, Lemke N. Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review. *Front Physiol* 2016;**7**:75.
126. Zhong J, Wang J, Peng W, *et al*. A feature selection method for prediction essential protein. *Tsinghua Sci Technol* 2015;**20**:491–9.
127. Guo F-B, Dong C, Hua H-L, *et al*. Accurate prediction of human essential genes using only nucleotide composition and association information. *Bioinformatics* 2017;**33**:1758–64.
128. Gustafson AM, Snitkin ES, Parker SC, *et al*. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 2006;**7**:265.
129. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* 2009;**10**:290.
130. Zeng M, Li M, Fei ZH, *et al*. DeepEP: a deep learning framework for identifying essential proteins. In: *IEEE International Conference on Bioinformatics and Biomedicine*, *2018*, pp. 583–8. IEEE, Piscataway, NJ.
131. Chen S-H, Chin C-H, Wu H-H, *et al*. cyto-Hubba: a Cytoscape plug-in for hub object analysis in network biology. *20th International Conference on Genome Informatics*, *2009*. Imperial College Pr.
132. Tang Y, Li M, Wang J, *et al*. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems* 2015;**127**:67–72.
133. Gräßler J, Koschützki D, Schreiber F. CentiLib: comprehensive analysis and exploration of network centralities. *Bioinformatics* 2012;**28**:1178–9.

134. Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. *Bioinformatics* 2009;**25**:2857–9.
135. Assenov Y, Ramírez F, Schelhorn S-E, *et al*. Computing topological parameters of biological networks. *Bioinformatics* 2007;**24**:282–4.
136. Konganti K, Wang G, Yang E, *et al*. SBEToolbox: a Matlab toolbox for biological network analysis. *Evol Bioinform Online* 2013;**9**:355–62.
137. Drozdov I, Ouzounis CA, Shah AM, *et al*. Functional Genomics Assistant (FUGA): a toolbox for the analysis of complex biological networks. *BMC Res Notes* 2011;**4**:462.
138. Junker BH, Koschützki D, Schreiber F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* 2006;**7**:219.
139. Li M, Yang J, Wu F-X, *et al*. DyNetViewer: a Cytoscape app for dynamic network construction, analysis and visualization. *Bioinformatics* 2017;**34**:1597–9.
140. Von Mering C, Krause R, Snel B, *et al*. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 2002;**417**:399–403.
141. Lawson CE, Wu S, Bhattacharjee AS, *et al*. Metabolic network analysis reveals microbial community interactions in anammox granules. *Nat Commun* 2017;**8**:15416.
142. Li M, Zheng R, Li Y, *et al*. MGT-SM: a method for constructing cellular signal transduction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2017. 10.1109/TCBB.2017.2705143.
143. Zheng R, Li M, Chen X, *et al*. BiXGBoost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics* 2018. 10.1093/bioinformatics/bty908.
144. Chen X, Li M, Wu F-X, *et al*. A novel method of gene regulatory network structure inference from gene knock-out expression data. *Tsinghua Sci Technol* 2018. 10.26599/TST.2018.9010097.
145. Yang L, Wang S, Zhou M, *et al*. Characterize the relationship between essential and TATA-containing genes for S. cerevisiae by network topologies in the perturbation sensitivity network. *Genomics* 2016;**108**:177–83.
146. Han HW, Ohn JH, Moon J, *et al*. Yin and Yang of disease genes and death genes between reciprocally scale-free biological networks. *Nucleic Acids Res* 2013;**41**:9209–17.
147. Doerr A. Single-cell proteomics. *Nat Methods* 2019;**16**:20.
148. Zeng M, Li M, Fei Z, *et al*. A deep learning framework for identifying essential proteins by integrating multiple sources of biological information. *IEEE/ACM Trans Comput Biol Bioinform 2019*. 10.1109/TCBB.2019.2897679.
149. Zhong J, Wang J, Peng W, *et al*. Prediction of essential proteins based on gene expression programming. *BMC Genomics* 2013;**14**:S7.
150. Fraser HB, Hirsh AE, Steinmetz LM, *et al*. Evolutionary rate in the protein interaction network. *Science* 2002;**296**:750–2.
151. Jordan IK, Rogozin IB, Wolf YI, *et al*. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 2002;**12**:962–8.
152. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol* 2006;**2**:e88.