Statistical Identification of Important Nodes in Biological Systems*

WANG Pei

DOI: 10.1007/s11424-020-0013-0
Received: 4 February 2020 / Revised: 23 May 2020

Abstract Biological systems can be modeled and described by biological networks. Biological networks are typical complex networks with widely real-world applications. Many problems arising in biological systems can be boiled down to the identification of important nodes. For example, biomedical researchers frequently need to identify important genes that potentially leaded to disease phenotypes in animal and explore crucial genes that were responsible for stress responsiveness in plants. To facilitate the identification of important nodes in biological systems, one needs to know network structures or behavioral data of nodes (such as gene expression data). If network topology was known, various centrality measures can be developed to solve the problem; while if only behavioral data of nodes were given, some sophisticated statistical methods can be employed. This paper reviewed some of the recent works on statistical identification of important nodes in biological systems from three aspects, that is, 1) in general complex networks based on complex networks theory and epidemic dynamic models; 2) in biological networks based on network motifs; and 3) in plants based on RNA-seq data. The identification of important nodes in a complex system can be seen as a mapping from the system to the ranking score vector of nodes, such mapping is not necessarily with explicit form. The three aspects reflected three typical approaches on ranking nodes in biological systems and can be integrated into one general framework. This paper also proposed some challenges and future works on the related topics. The associated investigations have potential real-world applications in the control of biological systems, network medicine and new variety cultivation of crops.

Keywords Biological network, complex network, important node, network motif, RNA-seq.

WANG Pei
School of Mathematics and Statistics, Institute of Applied Mathematics, Laboratory of Data Analysis Technology, Henan University, Kaifeng 475004, *China.* Email: wp0307@126.com; wangpei@henu.edu.cn.

*This paper was supported by the National Natural Science Foundation of China under Grant No. 61773153, the Natural Science Foundation of Henan under Grant No. 202300410045, the Supporting Plan for Scientific and Technological Innovative Talents in Universities of Henan Province under Grant No. 20HASTIT025, and the Training Plan of Young Key Teachers in Colleges and Universities of Henan Province under Grant No. 2018GGJS021. Partly supported by the Supporting Grant of Bioinformatics Center of Henan University under Grant No. 2018YLJC03.

◇ *This paper was recommended for publication by Editor GUO Jin.*

1 Introduction

Complex network science has interfused with many other scientific areas and has wider and wider real-world applications[1–6]. Plenty of real-world systems can be described or modeled by complex networks. Such as WWW, Internet, citation networks among scientific journals or authors, social systems and biological systems. Among which, biological systems are typical complex systems[1, 7–16].

Biological systems can be described by complex networks at different levels, generally including transcriptional regulatory networks (TRNs), gene regulatory networks (GRNs), protein-protein interaction (PPI) networks, metabolic networks, signaling networks, and so on[7]. Social networks are another typical example of complex networks. Spreading phenomena in social networks are ubiquitous, especially for epidemic spreading[17–20], such as SARS[21, 22], SARS-CoV-2[23–25]. Structure of social networks may affect disease spreading, and thus infectious diseases should be controlled via different approaches in different types of networks. Therefore, one should put the investigation of disease transmission and control in complex social networks. Before the complex network science was widely known, scientists mainly considered the spreading rules of diseases, seldom considered the topological structures of social networks. Network topological structures have strong effects on epidemic spreading[18–20]. Existing works reported that infectious disease can be easily spreading among people in scale-free networks[26–28]. Unfortunately, many real-world systems have the scale-free property, which indicates that it is actually difficult to control spreading phenomena in scale-free networks through control transmission strengths, unless spreaders were isolated[28].

Finding influential spreaders is the first step to control infectious disease spreading or design immunization strategies in social networks. Generally, nodes in complex systems are heterogeneous, indicating that nodes with different topological features or dynamical behaviors may have great differences on infection scopes, thus, a fundamental question is how to rank nodes in a complex system? Or in other word, how to identify important nodes? In fact, under different circumstances, important nodes have different meanings. From the epidemic perspective in complex networks, important nodes are equivalent to influential spreaders, nodes with high propagation capability/spreading scope[29], and so on. From the functional perspective in biological systems, important nodes may mean important genes[30], or disease genes[31], or functional genes[11–16, 32, 33]. In the following, we indiscriminately call important nodes as key nodes, crucial nodes, and so on.

Many problems in real-world systems can be boiled down to the identification of important nodes[19, 32, 33]. For example, the identification of influential spreaders in social networks, the identification of biomarkers or disease-causing genes in biological systems, the screening of stress responsive crucial genes in plants, and the selection of key points in electrical systems. To facilitate the identification of important nodes in biological systems, one needs to know either network structures among entities in the systems or behavioral data of entities. Once we obtained the networks or data from the biological systems, one of our ultimate goals was to identify the most important entities in the system.

A complex network consists of nodes and edges[1]. Mathematically, it can be described by an adjacency matrix $\boldsymbol{A} = (a_{ij})_{n\times n}$. Here, n denotes the number of nodes in the complex network. The elements of the adjacency matrix $\boldsymbol{A}$ are always non-negative, $a_{ij} > 0$ if node i and node j were connected, otherwise, $a_{ij} = 0$. Moreover, if $a_{ij} = a_{ji}$ for any node i and j, we call network $\boldsymbol{A}$ is undirected, otherwise, it is directed; If a_{ij} took either 0 or 1, then it is called an unweighted network; Otherwise, it is weighted. Mathematically, node ranking in a complex network is equivalent to find a mapping[29] from the network $\boldsymbol{A} = (a_{ij})_{n\times n}$ to node importance vector $S = (s_1, s_2, \cdots, s_n)^{\mathrm{T}}$:

$$F : \boldsymbol{A} \longrightarrow S. \tag{1}$$

Here, s_i represents the importance of node i in the network. Additionally, if there were some further node attributes (such as functional annotations of genes, gene expression profiles in different samples) described by matrix $\boldsymbol{B} = (b_{ij})_{n\times d}$ and edge attributes (such as repression or activation) described by matrix $\boldsymbol{C} = (c_{ij})_{n\times h}$, then the mapping may be written as:

$$F : f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \longrightarrow S. \tag{2}$$

Here, d represents the maximum number of node attributes, n represents the total number of nodes in the network, and h denotes the maximum degree of nodes. c_{ij} represents the value of edge attribute for the j'th edge of node i. $f(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ was a function of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. Actually, the problem can be degenerated into a network embedding problem or dimensional reduction problem[34, 35]. It is noted that the mapping F or function $f(\cdot)$ can be with explicit form, but usually, it is difficult to find the explicit form, such as those methods based on the integration of matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

With the rapid development of complex network science, various methods have been established to rank nodes if topological structure of complex networks were known[19]. These methods include the neighbor-based methods, path-based methods, iterative refinement measures and so on. But different methods have different advantages and different scopes of applications; There are no universal methods for all cases. Therefore, new measures are continuously proposed. Especially, with the rapid development of big data, some sophisticated statistical methods[36] have been proposed to deal with the cases where only behavioral data of entities in the complex system were known. In this paper, we introduce some works on the statistical identification of important nodes from three aspects, including our works in general complex networks[29], in biological networks[30, 37] and based on RNA-seq data[32]. The three approaches depend on different conditions, which can be integrated into one general framework, and seen as the dimensional reduction problems. The rest paper is organized as follows. From Sections 2 to 4, we introduce the three aspects of works, and in the last Section 5, we propose some challenge problems of the related topics and give some concluding remarks.

2 Statistical Identification of Important Nodes in General Complex Networks

2.1 The SpectralRank and Weighted SpectralRank Algorithms

Generally, the importance of a node in a complex network can be reflected by the quantity and quality of its neighbors, as well as the shortest paths that pass from this node. Thus, many neighbor-based methods, path-based methods and iterative refinement measures have been proposed[19]. Neighbor-based methods include degree centrality, semi-local centrality, k-shell, h-index, ING[38]; path-based methods include betweenness centrality, eccentricity, closeness, Katz centrality; PageRank (PR)[39], LeaderRank[40, 41] (LR), and eigenvector centrality are iterative refinement measures[19]. The mentioned measures have broad applications in real-world systems. For example, the well-known PR algorithm was originally used by Google to rank websites.

Each measure has limited scopes of applications and many of them can not be well explained through statistical theory[29]. For example, for the PR and LR algorithms in a degree uncorrelated network, based on mean-field theory, we proved that the average importance score for nodes within the node group with degree $k = (k^{out}, k^{in})$ is proportional to k^{in}, that is,

$$s_{(k)} \approx \theta k^{in}, \quad \theta = n/[(n+1)\langle k^{in}\rangle]. \tag{3}$$

Here, $s_{(k)}$ is the average importance score for nodes within the node group with degree $k = (k^{out}, k^{in})$. k^{out}, k^{in} denote out-degree and in-degree respectively. n denotes the number of nodes in the complex network. $\langle k^{in}\rangle$ denotes the average in-degree of the network.

Similar to the LR, by adding a ground node that was bidirectionally connected all the n nodes in the network, we propose two novel algorithms to rank node's propagation capability, which are called Spectralrank (SR) and weighted Spectralrank (WSR)[29]. Mathematically, the algorithms are described as:

$$S = c\widetilde{\boldsymbol{A}}S, \tag{4}$$

$$S_w = c\boldsymbol{W}S_w. \tag{5}$$

Here, S denotes the SR score and S_w denotes the WSR score; $\widetilde{\boldsymbol{A}}$ is the adjacency matrix of the augmented network,

$$\widetilde{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{A} & 1_{\boldsymbol{n}} \\ \boldsymbol{1}_n^{\mathrm{T}} & 0 \end{pmatrix}. \tag{6}$$

$\boldsymbol{W} = \widetilde{\boldsymbol{A}} + P$, $P = \operatorname{diag}\{p_1, p_2, \cdots, p_n, p_{n+1}\}$ is a priori knowledge of node's importance score, and we always set $p_{n+1} = 0$ for the ground node. c is a tuning parameter and is usually selected as the reciprocal of dominant eigenvalue of $\widetilde{\boldsymbol{A}}$ or $\boldsymbol{W}$, that is $c = 1/\lambda_1$. On the basis of this fact, the SR and WSR scores reduce to the dominant eigenvectors of $\widetilde{\boldsymbol{A}}$ and $\boldsymbol{W}$, respectively.

The algorithms include three steps: Firstly, we add a ground node to the graph, which bidirectionally connected with all other nodes in the complex network. This step can make all nodes in the network strongly connected. In the second step, if the WSR was considered, we add the a priori knowledge, which is a diagonal matrix, with values represent a priori knowledge of node importance. If SR was considered, we omit the second step. We calculate the principle eigenvector of the obtained matrix as SR or WSR score in the last step. Figure 1 shows an illustrative example. If we consider the total degree of each node in the original network as a priori knowledge, we obtain the following weighted augmented adjacency matrix:

$$\boldsymbol{W}=\begin{pmatrix}2&1&0&0&1&0&0&1\\1&3&0&0&0&0&0&1\\0&1&2&1&0&0&0&1\\0&0&1&2&0&0&0&1\\1&0&0&1&3&1&0&1\\0&0&0&0&0&1&0&1\\0&1&0&0&0&0&1&1\\1&1&1&1&1&1&1&0\end{pmatrix}. \tag{7}$$

The dominant eigenvectors of $\widetilde{\boldsymbol{A}}$ and W can be obtained as follows:

$$S=(0.3676,0.2774,0.3280,0.2660,0.4034,0.1715,0.2514,0.5951)^{\mathrm{T}},$$

$$S_w=(0.4207,0.3843,0.3260,0.2343,0.5370,0.1008,0.1924,0.4226)^{\mathrm{T}}.$$

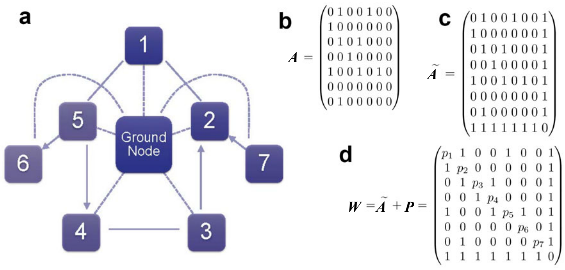
Therefore, the ranks for the seven nodes in the original network according to the SR is

$$\text{Rank}_{SR}=(2,4,3,5,1,7,6)^{\mathrm{T}}$$

and

$$\text{Rank}_{WSR}=(2,3,4,5,1,7,6)^{\mathrm{T}}$$

corresponds to the WSR. Obviously, the rankings from the SR and WSR were quite similar, except for nodes 2 and 3. This is because that the total degree for node 2 is 3, which is larger than that for node 3, thus, if this priori knowledge was considered, the WSR ranks node 2 more important than node 3.



$$\boldsymbol{A}=\begin{pmatrix}0&1&0&0&1&0&0\\1&0&0&0&0&0&0\\0&1&0&1&0&0&0\\0&0&1&0&0&0&0\\1&0&0&1&0&1&0\\0&0&0&0&0&0&0\\0&1&0&0&0&0&0\end{pmatrix}$$

$$\widetilde{\boldsymbol{A}}=\begin{pmatrix}0&1&0&0&1&0&0&1\\1&0&0&0&0&0&0&1\\0&1&0&1&0&0&0&1\\0&0&1&0&0&0&0&1\\1&0&0&1&0&1&0&1\\0&0&0&0&0&0&0&1\\0&1&0&0&0&0&0&1\\1&1&1&1&1&1&1&0\end{pmatrix}$$

$$\boldsymbol{W}=\widetilde{\boldsymbol{A}}+\boldsymbol{P}=\begin{pmatrix}p_1&1&0&0&1&0&0&1\\1&p_2&0&0&0&0&0&1\\0&1&p_3&1&0&0&0&1\\0&0&1&p_4&0&0&0&1\\1&0&0&1&p_5&1&0&1\\0&0&0&0&0&p_6&0&1\\0&1&0&0&0&0&p_7&1\\1&1&1&1&1&1&1&0\end{pmatrix}$$

Figure 1 An illustrative example[29]. (a) An augmented network with a ground node. (b) The adjacency matrix of the original network. (c) The adjacency matrix of the augmented network as shown in (a). (d) The augmented adjacency matrix with a priori knowledge

2.2 Statistical Explanation and Applications in Biological Networks

The algorithms are very simple and efficient. Moreover, we established some probability frameworks to illustrate that the proposed algorithms are statistically meaningful[29]. For simplicity, we considered undirected and unweighted complex networks, and we assumed that the networks follow the fitness growth model, that is, the probability $p(i,j)$ of adding an edge between nodes i and j is proportional to the product of their importance scores s_i and s_j:

$$p(i,j) \propto s_i s_j. \tag{8}$$

Furthermore, we supposed that the generated networks follow the Boltzmann distribution[42]:

$$p(\boldsymbol{A};\boldsymbol{s}) = \frac{\mathrm{e}^{-H(\boldsymbol{A};\boldsymbol{s})}}{Z_{\mathcal{A}}}, \tag{9}$$

where the energy is given by the Hamiltonian function $H(\boldsymbol{A};\boldsymbol{s}) = -\sum_{i,j\in V} a_{ij}s_is_j$ and $Z_{\mathcal{A}} = \sum_{\boldsymbol{A}\in\mathcal{A}} p(\boldsymbol{A};\boldsymbol{s})$ is the partition constant. Note that this distribution coincides with the Ising model without an external field[42]. $\mathcal{A}$ denotes the ensemble of all the generated networks.

Based on the above assumptions, the following conclusions were obtained:

1) The maximum likelihood estimation of importance score vector S in the fitness model is exactly the eigenvector centrality of the network $\boldsymbol{A}$ under the constraint $S^{\mathrm{T}}S = 1$. Furthermore, $c = 1/\lambda_1$ is the necessary condition of the maximum likelihood estimation:

$$S = \arg\max_{\boldsymbol{s}}\ \log p(\boldsymbol{A};\boldsymbol{s}) = \arg\max_{\boldsymbol{s}} \sum_{i,j\in V} a_{ij}s_is_j = \arg\max_{\boldsymbol{s}}\ \boldsymbol{s}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{s}. \tag{10}$$

Here, $s = (s_1, s_2, \cdots, s_{n+1})^{\mathrm{T}}$ and $s^{\mathrm{T}}s = 1$, similarly hereinafter.

2) We assume that a priori distribution of s is governed by the conjugate a priori $p(s) = \mathrm{e}^{-s^{\mathrm{T}}Ps}/Z_F$, $Z_F = \int p(s)ds$ is a partition constant. We deduced that the priori knowledge P functioned as a L_2 norm penalty, which can effectively prevent over-fitting:

$$\begin{aligned} S &= \arg\max_{\boldsymbol{s}}\ \log p(\boldsymbol{s}|\boldsymbol{A}) = \arg\max_{\boldsymbol{s}}\ \log p(\boldsymbol{A}|\boldsymbol{s}) + \log p(\boldsymbol{s}) \\ &= \arg\max_{\boldsymbol{s}}\ \boldsymbol{s}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{s} + \sum_{i=1}^{N} s_i^2\boldsymbol{P}(i,i). \end{aligned} \tag{11}$$

3) The addition of the ground node is also a kind of L_2 norm penalty, which can also prevent over-fitting:

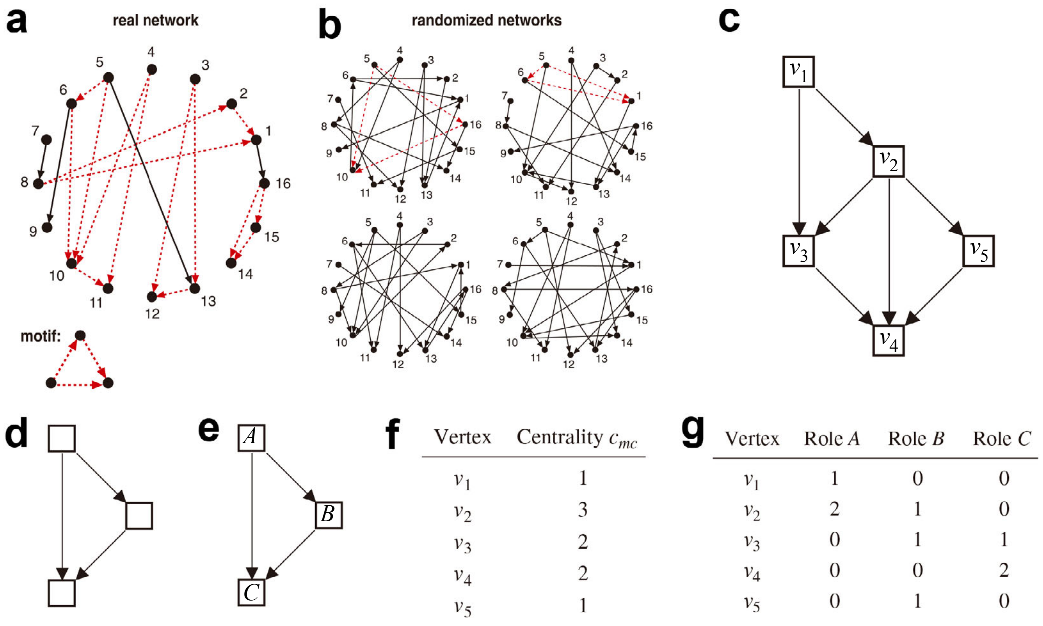
$$S = \arg\max_{\boldsymbol{s}}\ \boldsymbol{s}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{s}^{\mathrm{T}}\left(\frac{2}{\lambda}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{1}\right)\boldsymbol{s}. \tag{12}$$

For details, one can refer to [29].

The proposed algorithms were evaluated by 32 real-world networks[29]. The actual importance scores of nodes were simulated according to the SIR model. Each node was successively set as initial spreader, and the final propagation scope of a node after the spreading process reached its stable state was taken as its actual importance score. The actual average importance score vector was obtained by averaging over 100 independent simulation runs. The

Kendall correlation coefficient between actual importance score vector and S was used to evaluate the performance of the proposed algorithm. Numerical results show that the proposed algorithms outperform many other existing algorithms, and it can be applied to binary networks, undirected and directed networks.

We also applied the algorithms to biological networks, including the neural network of C. elegans, TRN of E. coli. Taken the command interneurons in the neural network, the 18 global regulators and 7 key global regulators in the TRN as gold standards, we performed ROC analysis, results show that the proposed algorithms also have good performance in biological networks. The applicability of the algorithms in biological networks indicated that functional important transcription factors (TFs) or neurons may also play important roles in signaling spreading.



Vertex	Centrality c_{mc}
v_1	1
v_2	3
v_3	2
v_4	2
v_5	1

Vertex	Role A	Role B	Role C
v_1	1	0	0
v_2	2	1	0
v_3	0	1	1
v_4	0	0	2
v_5	0	1	0

Figure 2 Network motif[43] and motif centrality[44]. Network motifs are patterns that recur much more frequently in the real network (a) than in an ensemble of randomized networks (b). Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. Red dashed lines indicate edges that participate in the FFL motif, which occurs five times in the real network. (c) A target graph; (d) the FFL motif; (e) The FFL motif with three different roles A, B and C. (f) The motif-based centrality given by the FFL without roles; (g) The extended motif-based centrality given by the FFL with roles

3 Statistical Identification of Important Nodes in Biological Networks

3.1 Network Motifs

Real-world biological networks are too complex, which have hindered our comprehensive understanding of them. Fortunately, researchers have founded that biological networks consist of simple regulatory circuits, called as network motifs[43]. Thus, investigations on network motifs are the first step to the system level understanding of biological networks.

The concept of network motifs was proposed by Uri Alon and coauthors in the year 2002[43]. Network motifs are patterns of interconnections occurring in a complex network at numbers that are significantly higher than those in randomized networks (Figures 2 a and b). For each network, one generates hundreds of randomized networks. The number of a subgraph in the real-world network is denoted as N_{rl}. The average number in random networks is denoted as N_{rd}, with standard deviation denoted by S_d. The Z_{score} measures the significance of the subgraph[43], which is defined as $Z_{\text{score}} = (N_{rl} - N_{rd})/S_d$. Another index U is defined as the number of times a subgraph appears in the investigated network with distinct sets of nodes. Generally, subgraphs with $Z_{\text{score}} \geq 2, U \geq 4$ and $N_{rl} - N_{rd} \geq 0.1N_{rd}$ are identified as motifs.

Uri Alon and coauthors reported that the three-node feed-forward loop (FFL) is a typical network motif. FFLs thus attracted wide attentions. Many works have been published to clarify the relationships among structures, functions and dynamics of network motifs[45].

Network motifs are building blocks of complex biological networks, thus, they can be used to evaluate node importance in biological networks. Previously, some researchers have considered to use network motifs to rank nodes. For example, Koschützki, et al.[44] proposed a node centrality measure based on the FFL and the role of each node in the FFL (Figures 2 c–g). They counted the frequency of each node involved in the FFL according to three roles, and nodes can be ranked by the sum of their total frequencies or according to each role. Except that, network motifs have been also used to evaluate node importance in neuron networks[46–49].

3.2 A Novel Network Motif Centrality Measure for Directed Biological Networks

Motivated by existing works and network motifs, we proposed a new measure based on principal component analysis (PCA)[50] for directed biological networks[30]. The proposed method includes two steps: Motif counts and PCA analysis. Firstly, we detected all two-node, three-node and four-node network motifs in a biological network, and we counted the frequency of each node involved in each type of motifs, and then different motifs were weighted according to their total frequencies. Finally, we performed PCA analysis on the obtained data and our new measure was proposed as the first principal component.

Mathematically, suppose that there were totally m types of two-node, three-node and four-node network motifs. We denoted the occurrences of node i in the j-th type of motif as $u_{ij}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, m$. Then, we derived a matrix $U = (u_{ij})_{n\times m}$ for the network. In real-world networks, the importance of different types of motifs were varied. Therefore, we endowed each motif with a weight $w_j, j = 1, 2, \cdots, m$, where $w_j = c_j / \sum_{k=1}^{m} c_k$. Here, c_k $(k = 1, 2, \cdots, m)$ denotes the number of the k-th type of motif. Subsequently, we derived a revised matrix $B = (b_{ij})_{n\times m} = (b_1, b_2, \cdots, b_m) = (w_j u_{ij})_{n\times m}$. Based on B and the idea of the

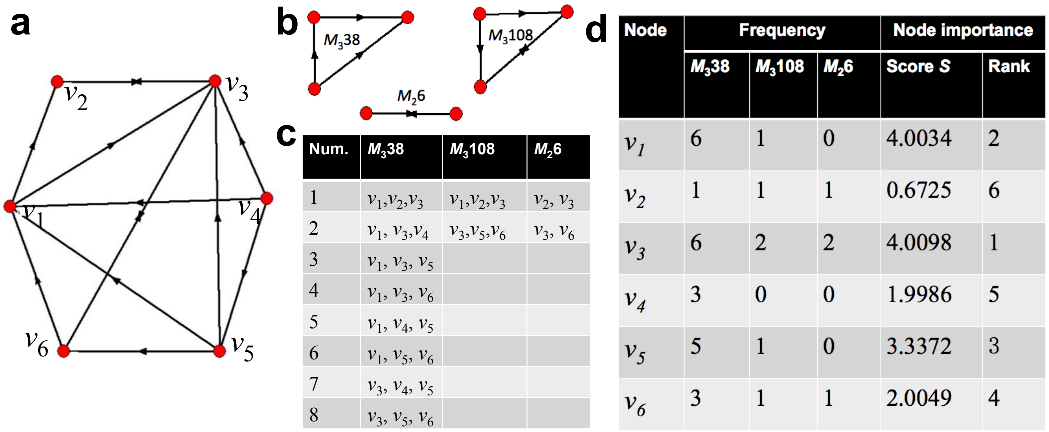
PCA[50], we constructed the following index to obtain importance score S for the n nodes in the network:

$$S = \sum_{j=1}^{m} \alpha_j b_j. \tag{13}$$

It was proved that the parameter vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)^{\mathrm{T}}$ was just the eigenvector of the dominant eigenvalue of the covariance matrix for matrix B[30]:

$$\Sigma = \frac{1}{n-1}(B^{\mathrm{T}}B - n\overline{B}\,\overline{B}^{\mathrm{T}}).$$

Here, $\overline{B}$ is the column mean vector of matrix B. Figure 3 shows an illustrative example for the proposed motif centrality[30].



Num.	$M_3$38	$M_3$108	$M_2$6
1	v_1,v_2,v_3	v_1,v_2,v_3	v_2, v_3
2	v_1, v_3,v_4	v_3,v_5,v_6	v_3, v_6
3	v_1, v_3, v_5		
4	v_1, v_3, v_6		
5	v_1, v_4, v_5		
6	v_1, v_5, v_6		
7	v_3, v_4, v_5		
8	v_3, v_5, v_6		

Node	Frequency			Node importance	
	$M_3$38	$M_3$108	$M_2$6	Score S	Rank
v_1	6	1	0	4.0034	2
v_2	1	1	1	0.6725	6
v_3	6	2	2	4.0098	1
v_4	3	0	0	1.9986	5
v_5	5	1	0	3.3372	3
v_6	3	1	1	2.0049	4

Figure 3 An illustrative example for the proposed motif centrality[30]. (a) A simple network with six nodes. (b) Subgraphs that are assumed to be motifs in network (a). (c) Members that consist of the three types of motifs. (d) Appearances of nodes in each motif and node importance rankings according to the proposed motif centrality

We applied the proposed algorithm to five real-world biological networks, including the neuronal network for C. elegan, TRNs for E. coli, drosophila and yeast, a signal transduction network for human[30]. These networks encompassed hundreds to thousands of nodes, and tens to tens of thousands of network motifs. ROC curve analysis revealed that the proposed method can well identify command interneurons in the neuron network, and global regulators in the TRN for E. coli. The proposed algorithm can exclude not important hubs but rank non-hub and actually important nodes at the top. Moreover, based on rich-club analysis, the proposed algorithm can also help us to find densely connected clusters, the top ranked nodes were more densely connected than those identified by the other methods.

3.3 An Integrative Measure for Undirected PPI Network: An Extension of the Motif Centrality

The above work only considered directed biological networks. PPI networks are generally undirected. Extensive measures have been proposed to evaluate the structural importance of a node in a complex network[19]. Motivated by the proposed motif centrality and based on PCA, by integrating different centrality measures via PCA, we proposed an integrative measure in PPI networks, which can help to identify structural dominant proteins (SDPs)[37]. The proposed measures can integrate many existing measures (Figure 4). For simplicity, we considered the well-known degree centrality, betweenness, closeness, k-shell, semi-local centrality and motif centrality measures, and integrated them together.

Based on literature survey and existing databases, we constructed several real-world PPI networks with different sizes for the yeast. Moreover, in order to see the evolution of SDPs, we also constructed artificial PPIs, which are based on the duplication-divergence (DD) model[37]. We considered the anti-preference duplication process and the edge deletion, dimerization, edge addition and isolated node removal divergence processes. The DD model can generate an ensemble of random networks. The artificial PPI networks have similar topological features as the real-world ones with fine tuning parameters in the DD model[37].

By applying the algorithm to the constructed PPI networks, we can find SDPs in the PPI networks. Moreover, we found that only a small fraction of proteins were structurally dominant. SDPs evolved more slowly than unimportant nodes. Targeted mutations on SDPs can keep certain robustness, as compared with targeted mutations on hubs[37].

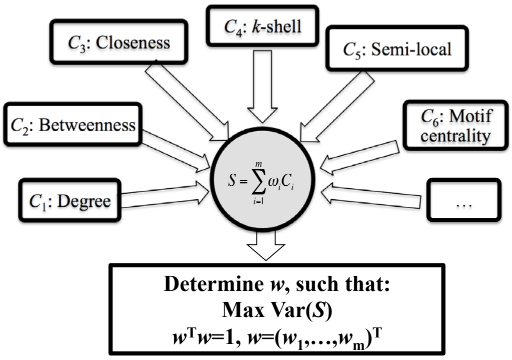


Figure 4 Basic idea of the integrative measure for undirected PPI network

4 Statistical Identification of Important Genes Based on RNA-Seq Data

4.1 RNA-Seq

RNA sequencing (RNA-seq) uses the next generation sequencing (NGS) technologies to reveal the presence and quantity of RNA molecules in biological samples. RNA-seq analyses can be performed at four different levels: Sample-level, gene-level, transcript-level, and exon-level[51]. In the sample-level analysis, the results are usually summarized into a similarity matrix. The gene-level analysis summarizes the counts of RNA-seq reads mapped to genes in samples of different conditions, and it subsequently compares genes' expression levels that were calculated based on read counts. The transcript-level analysis focuses on reads mapped to different isoforms. The exon-level analysis mostly considers the reads mapped to or skipping the exon of interest.

A flow chart of RNA-seq analysis typically includes the following steps: Experimental design, RNA sequencing, data analysis, biological mechanism clarification and experimental verification (Figure 5). The first step is to design experiments and cultivate samples for certain purpose. And then sequencing the samples via high-throughput sequencer and obtaining sequence data. The subsequently complicated work is to perform data analysis. Such as sequence alignment, finding differentially expressed genes (DEGs) and perform hypothesis test. One of the ultimate goal of data analysis is to find important genes that cause phenotype variation among experimental samples. Based on the selected genes, researchers can clarify the related molecular mechanism. The obtained results should be verified according to qRT-PCR experiments or functional verification (such as gene mutation experiments) for further applications in the cultivation of new crop varieties[10, 32, 33].

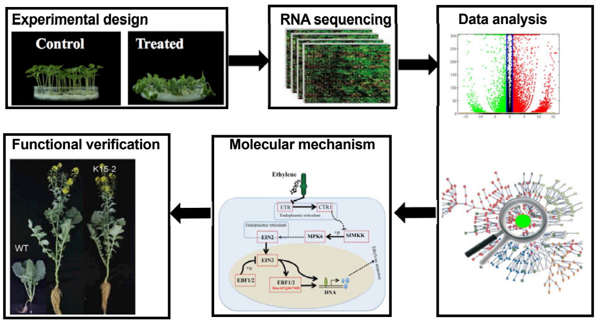


Figure 5 A flow chart of RNA-seq analysis typically includes the following steps: Experimental design, RNA sequencing, data analysis, biological mechanism clarification and experimental verification

4.2 Important Genes Identification Based on RNA-Seq Data

As we have mentioned, one of the most important goal of RNA-seq data analysis is the identification of important genes that possibly cause phenotype variation. Many methods have been proposed to cope with this problem. Traditionally, $\log_2$ fold change values and hypothesis tests are frequently used to select crucial DEGs[32, 33, 51]. If there are m treated samples and m samples served as controls, and the expression values (FPKM, RPKM or other methods) of gene i $(i = 1, 2, \cdots, n)$ in the treated samples are denoted as $y_i = (y_{i1}, y_{i2}, \cdots, y_{im})^{\mathrm{T}}$, and those in the controlled samples are denoted as $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})^{\mathrm{T}}$. The mean expression values under treatment and control are denoted as $\overline{y}_i$ and $\overline{x}_i$ respectively. Then the $\log_2$ fold change value for gene i was defined as:

$$\log_2(FC_i) = \log_2(\overline{y}_i/\overline{x}_i). \tag{14}$$

$\log_2(FC_i) > 1$ indicates that gene i was up-regulated under treatment in comparison with control, and the expression under treatment was more than two times higher than that under control; $\log_2(FC_i) < -1$ indicates that gene i was down-regulated under treatment in comparison with control, and the expression under control was more than two times higher than that under treatment.

Moreover, one can perform hypothesis t test (the expression values of genes in repeated samples under a certain condition are assumed to be normally distributed) to verify whether the expression of gene i under treatment was significantly different from that under control. Generally, due to experimental costs, the sample size m is very small. Thus, the revised t statistic can be used to test the null hypothesis that the expression values of a gene between the treated and the control samples have no significant difference, which is written as[52]:

$$t_i = \frac{\overline{y}_i - \overline{x}_i}{\sqrt{s_{yi}^2/m + s_{xi}^2/m}}. \tag{15}$$

Here, $s_{yi}^2 = \sum_{j=1}^m (y_{ij} - \overline{y}_i)^2/(m-1), s_{xi}^2 = \sum_{j=1}^m (x_{ij} - \overline{x}_i)^2/(m-1)$. Under the null hypothesis, the t_i statistic follows the t distribution with degree of freedom:

$$df_i = \frac{(m-1)(s_{yi}^2 + s_{xi}^2)^2}{s_{yi}^4 + s_{xi}^4}.$$

Conventionally, one can use $|\log_2(FC_i)| > 1$ and $P_i = p\{|t(df_i)| > t_i\} < 0.05$ as criterion to evaluate whether gene i was a DEG. If there are too many DEGs, one can choose larger threshold values for $\log_2(FC)$ and P. For example, $|\log_2(FC_i)| > 2$ and $P < 0.01$. Recently, some novel methods to screen DEGs from RNA-seq data have been proposed, which were based on the assumption that the read counts follow negative binomial distribution, and the related R package DESeq[53] and DESeq2[54] were successively developed.

Except the above mentioned method, researchers have developed many other methods to identify crucial genes from omics data. For example, Li, et al.[55], Chen, et al.[56] and Hou, et al.[57] introduced the hidden Markov model (HMM) for genome-wide association studies (GWAS). Among which, based on the HMM and the 'guilt-by-rewiring' principle of the gene co-expression networks, Hou, et al.[57] identified disease genes in crohn's disease and parkinson's disease.

In the following, taking one of our works on Brassica napus (B. napus) as an example, we briefly introduce one work on the identification of important genes based on RNA-seq data[32]. The Brassica genus includes a diverse range of vegetable and oilseed crops which are important for human nutrition, such as Brassica rapa (B. rapa), Brassica oleracea (B. oleracea) and B. napus. The whole genome map for B. napus[58], B. rapa[59], and B. olercea[60] and Arabidopsis thaliana[61] have been published, which facilitate us to genome-widely explore these species.

Plants encompass diverse TF families, such as AP2, bZIP, MYB, NAC and WRKY[32]. Some TF families are crucial for stress responsiveness in plants. Thus, the identification of TF families in B. napus is an interesting yet important problem. Through gene and protein sequence alignment, we identified five families of TFs in the four species. We genome-widely identified totally 2167 TFs in B. napus belonging to the five families, including 518 BnAP2/EREBPs, 252 BnbZIPs, 721 BnMYBs, 398 BnNACs and 278 BnWRKYs, which contained some novel members in comparison with existing results[32].

We performed structural analysis, synteny analysis and cis-acting element analysis on the identified TFs[32]. Sub-genome distributions of BnAP2/EREBPs and BnMYBs indicated that the two families might have suffered from duplication and divergence during evolution. Phylogenetic analysis revealed that each TF family can be divided into several subfamilies according to their sequence similarity. Synteny analysis revealed strong co-linearity between B. napus and its two ancestors, although chromosomal rearrangements have occurred and 85 TFs were lost. About 7.6% and 9.4% TFs of the five families in B. napus were novel genes and conserved genes, which both showed preference on the C sub-genome.

To see the responsiveness of the five TF families under stress, we designed RNA-seq experiments and obtained RNA-seq data. We cultivated B. napus seedings of 7-day-old as samples, and we considered five treatments, including cold, heat, drought, salt and ABA, The seedlings without any treatments were taken as controls. Seedlings were sampled at 12h after treatments for RNA extraction. Each experiment was repeated three times. The samples were sequenced by using the Illumina HiSeq 4000 platform. RNA-Seq data reveals that 449 of the 518 BnAP2/EREBPs, 227 of the 252 BnbZIPs, 585 of the 721 BnMYBs, 332 of the 398 BnNACs, 241 of the 278 BnWRKYs respond to at least one of the five treatments. Based on $\log_2(FC)$ (based on FPKM values), hypothesis test and GO annotations, totally 315 crucial DEGs were screened out, including 93 BnAP2/EREBPs, 42 BnbZIPs, 94 BnMYBs, 48 BnNACs and 38 BnWRKYs.

GO enrichment analysis revealed that the 315 DEGs totally enriched in 213 biological process terms ($P < 0.01$), including various biological regulation processes, responding to various stimulus and diverse signaling pathways. Clustering analysis on expression values of the 315 DEGs revealed that crucial TFs in each family were hierarchically clustered[32]. The expression profiles of crucial TFs under drought, salt and ABA were all similar in the five families. TFs from the same subfamilies tended to be clustered.

For the 315 crucial TFs, based on gray correlation coefficient, we constructed gene co-expression networks, and performed comparative analysis with homologous gene network of A thalina. We found that the crucial TFs could trigger the differential expression of targeted genes, resulting in a complex clustered network with clusters of genes responsible for targeted stress responsiveness. To verify the reliability of the data, we verified 40 genes via qRT-PCR experiments. qRT-PCR results revealed that the obtained data are reproducible[32].

5 Discussions and Conclusions

In this paper, we mainly reviewed some recent works on the identification of important nodes in biological systems, which are based on three different approaches and depending on different conditions. We have proposed the SR and WSR algorithms to evaluate node propagation capability in general complex networks, and we build a new probabilistic framework, which provides a theoretic understanding on eigenvector centrality, ground node and a priori knowledge in the algorithm. The proposed SR and WSR can be seen as a mapping from the adjacency matrix $\boldsymbol{A}$ of a complex network to node importance score S. We also proposed a network motif centrality for directed biological networks, which is based on PCA and network motif detection; We further extended the idea of the motif centrality to undirected PPI networks, and we proposed an integrative measure, which can be used to identify SDPs in PPI networks. The motif centrality measure and the integrative measure can not be explicitly written as a mapping from the adjacency matrix $\boldsymbol{A}$ of a complex network. However, the motif participation matrix U or weighted matrix $\boldsymbol{B}$ is actually a function of the adjacency matrix $\boldsymbol{A}$, which can be described as $\boldsymbol{B} = g(\boldsymbol{A})$. Here, the function $g(\cdot)$ can not be explicitly written. In this case, the motif centrality can also be seen as a mapping from $\boldsymbol{A}$ to S, similarly for the integrative measure. We explored five TF families in B. napus and investigated their stress responsive characteristics, 315 crucial TFs were screened. The last work was totally based on data, thus, it is different from the previous two works. However, RNA-seq data are obtained from biological systems, which can also be constructed as a complex network. Thus, identifying crucial TFs from RNA-seq can be also seen as a mapping. It is interesting yet meaningful to establish a unified framework for the mentioned works, the recently developed graph representation theory[34] or graph neural networks[35] are promising tools.

Although the proposed methods have some advantages, there are still many issues to be further explored. For example, for the proposed motif centrality, if there are no network motifs in a biological system, then the algorithm will lose efficacy. However, one possible extension of the motif centrality is that one does not consider whether a subgraph was a network motif, and consider all two-node, three-node and four-node subgraphs. Another extension is that the PCA is a linear method, possibly one can extend the importance score S as a nonlinear function. However, nonlinear S will undoubtedly increase the computational difficulty. For the identification of important genes based on RNA-seq data, we mainly used the $\log_2$ fold change value and hypothesis test methods to identify crucial DEGs. Our future works will consider the reconstruction of gene co-expression networks based on RNA-seq data, and then we will further use the gene co-expression networks to identify crucial responsive genes[62]. A big challenge in RNA-seq data analysis is in that one often encounters the cases with $p >> n$, that is, the number of variables or features is far larger than the number of samples[63]. When $p >> n$, many traditional statistical methods lose their effectiveness, one must establish new methods to perform data analysis. Our ongoing works will consider the logistic regression method with various penalizations[63] to cope with such problem.

The associated works can help us to understand the complex biological systems, and they may have potential applications in biological network control, network medicine and new variety cultivation of crops.

References

[1] Newman M, Barabási A L, and Watts D J, *The Structure and Dynamics of Networks*, Princeton University Press, Princeton and Oxford, 2006.

[2] Wu X, Wei W, Tang L, et al., Coreness and h-index for weighted networks, *IEEE Trans. Circuits Syst. I: Reg. Papers*, 2019, **66**(8): 3113–3122.

[3] Mei G, Wu X, Wang Y, et al., Compressive-sensing-based structure identification for multilayer networks, *IEEE Trans. Cyber.*, 2018, **48**(2): 754–764.

[4] Wei X, Wu X, Chen S, et al., Cooperative epidemic spreading on a two-layered interconnected network, *SIAM J. Appl. Dyn. Syst.*, 2018, **17**(2): 1503–1520.

[5] Jia Z, Chen H, Tu L, et al., Stability and feedback control for a coupled hematopoiesis nonlinear system, *Adv. Differ. Equa.*, 2018, **2018**: 401.

[6] Long Y, Jia Z, and Wang Y, Coarse graining method based on generalized degree in complex network, *Physica A*, 2018, **505**: 655–665.

[7] Chen L, Wang R, and Zhang X, *Biomolecular Networks: Methods and Applications in Systems Biology*, Wiley, New Jersey, 2009.

[8] Liu S, Xu Q, Chen A, et al., Structural controllability of static and dynamic transcriptional regulatory networks for Saccharomyces cerevisiae, *Physica A*, 2020, **537**: 122772.

[9] Barabási A L, Gulbahce N, and Loscalzo J, Network medicine: A network-based approach to human disease, *Nat. Rev.*, 2011, **12**: 56–68.

[10] Wang Z, Yang C, Chen H, et al., Multi-gene co-transformation can improve comprehensive resistance to abiotic stresses in B. napus L., *Plant Sci.*, 2018, **274**: 410–419.

[11] Shang B, Zang Y, Zhao X, et al., Functional characterization of GhPHOT2 in chloroplast avoidance of Gossypium hirsutum, *Plant Physiol. Bioch.*, 2019, **135**: 51–60.

[12] Qu X, Cao B, Kang J, et al., Fine-tuning stomatal movement through small signaling peptides, *Front Plant Sci.*, 2019, **10**: 69.

[13] Wang D, Yang C, Dong L, et al., Comparative transcriptome analyses of drought-resistant and -susceptible Brassica napus L. and development of EST-SSR markers by RNA-Seq, *J. Plant Biol.*, 2015, **58**: 259–269.

[14] Zhang S, Li X, Pan J, et al., Use of comparative transcriptome analysis to identify candidate genes related to albinism in channel catfish (Ictalurus punctatus), *Aquaculture*, 2018, **500**: 75–81.

[15] Dong, W, Li M M, Li Z G, et al., Transcriptome analysis of the molecular mechanism of Chrysanthemum flower color change under short-day photoperiods, *Plant Physiol. Bioch.*, 2020, **146**: 315–328.

[16] Zhang G F, Yue C M, Lu T T, et al., Genome-wide identification and expression analysis of NADPH oxidase genes in response to ABA and abiotic stresses, and in fibre formation in Gossypium, *Peer J*, 2020, **8**: e8404.

[17] Kitsak M, Gallos L K, Havlin S, et al., Identification of influential spreaders in complex networks, *Nat. Phys.*, 2010, **6**: 888–893.

[18] Wang P, Tian C, and Lu J, Identifying influential spreaders in artificial complex networks, *Journal of Systems Science and Complexity*, 2014, **27**(4): 650–665.

[19] Lü L Y, Chen D, Ren X, et al., Vital nodes identification in complex networks, *Phys. Rep.*, 2016, **650**: 1–63.

[20] Zhang Z K, Liu C, Zhan X X, et al., Dynamics of information diffusion and its applications on complex networks, *Phys. Rep.*, 2016, **651**: 1–34.

[21] Ksiazek T G, Erdman D, Goldsmith C S, et al., A novel coronavirus associated with severe acute respiratory syndrome, *N. Engl. J. Med.*, 2003, **348**: 1953–1966.

[22] Kuiken T, Fouchier R, Schutten M, et al., Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome, *Lancet*, 2003, **362**: 263–270.

[23] Zhu N, Zhang D, Wang W, et al., A novel coronavirus from patients with pneumonia in China, *N. Engl. J. Med.*, 2020, **382**: 727–733.

[24] Huang C, Wang Y, Li X, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, *Lancet*, 2020, **395**: 497–506.

[25] Wang P, Lu J, Jin Y, et al., Statistical and network analysis of 1212 COVID-19 patients in Henan, China, *Int. J. Infect. Disease*, 2020, **95**: 391–398.

[26] Pastor-Satorras R and Vespignani A, Epidemic spreading in scale-free networks, *Phys. Rev. Lett.*, 2001, **86**(14): 3200–3203.

[27] Boguna M, Pastor-Satorras R, and Vespignani A, Absence of epidemic threshold in scale-free networks with degree correlations, *Phys. Rev. Lett.*, 2003, **90**(2): 028701.

[28] Gallos L K, Liljeros F, Argyrakis P, et al., Improving immunization strategies, *Phys. Rev. E*, 2007, **75**(4): 045104.

[29] Xu S, Wang P, Zhang C, et al., Spectral learning algorithm reveals propagation capability of complex network, *IEEE Trans. Cyber.*, 2019, **49**(12): 4253–4261.

[30] Wang P, Lü J, and Yu X, Identification of important nodes in directed biological networks: A network motif approach, *PLoS One*, 2014, **9**(8): e106132.

[31] Wang P, Chen Y, Lü J, et al., Graphical features of functional genes in human protein interaction network, *IEEE Trans. Biomed. Circuits Syst.*, 2016, **10**(3): 707–720.

[32] Wang P, Yang C, Chen H, et al., Exploring transcriptional factors reveals crucial members and regulatory networks involved in different abiotic stresses in Brassica napus L., *BMC Plant Biol.*, 2018, **18**: 202.

[33] Wang P, Yang C, Chen H, et al., Transcriptomic basis for drought-resistance in Brassica napus L., *Sci. Rep.*, 2017, **7**: 40532.

[34] Chen F, Wang Y, Wang B, et al., Graph representation learning: A survey, 2019, arXiv: 1909.00958.

[35] Wu Z, Pan S, Chen F, et al., A comprehensive survey on graph neural networks, 2019, ArXiv: 1901.00596v3.

[36] Bühlmann P and van de Geer S, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer-Verlag, Berlin Heidelberg, 2011.

[37] Wang P, Yu X, and Lü J, Identification and evolution of structurally dominant nodes in protein-protein interaction networks, *IEEE Trans. Biomed. Circuits Syst.*, 2014, **8**(1): 87–97.

[38] Xu S, Wang P, and Lü J, Iterative neighbour-information gathering for ranking nodes in complex networks, *Sci. Rep.*, 2017, **7**: 41321.

[39] Brin S and Page L, Reprint of: The anatomy of a large-scale hypertextual web search engine, *Comput. Netw.*, 2012, **56**(18): 3825–3833.

[40] Lü L, Zhang Y, Yeung C H, et al., Leaders in social networks, the delicious case, *PLoS One*, 2011, **6**: e21202.

[41] Xu S and Wang P, Identifying important nodes by adaptive LeaderRank, *Physica A*, 2017, **469**:

654–664.

[42] Metzner R, Fundamental of statistical and thermal physics, *Phys. Today*, 1967, **20**(12): 85–87.

[43] Milo R, Shen-Orr S, Itzkovitz S, et al., Network motifs: Simple building blocks of complex networks, *Science*, 2002, **298**: 824–827.

[44] Koschützki D, Schwöbbermeyer H, and Schreiber F, Ranking of network elements based on functional substructures, *J. Theor. Biol.*, 2007, **248**: 471–479.

[45] Alon U, Network motifs: Theory and experimental approaches, *Nat. Rev. Genet.*, 2007, **8**(6): 450–461.

[46] Koschützki D and Schreiber F, Centrality analysis methods for biological networks and their application to gene regulatory networks, *Gene Regulat. Syst. Biol.*, 2008, **2**: 193–201.

[47] Sporns O and Kötter R, Motifs in brain networks, *PLoS Biol.*, 2004, **2**: e369.

[48] Sporns O, Honey C J, and Kötter R, Identification and classification of hubs in brain networks, *PLoS One*, 2007, **2**: e1049.

[49] Rubinov M and Sporns O, Complex network measures of brain connectivity: Uses and interpretations, *NeuroImage*, 2010, **52**: 1059–1069.

[50] Härdle W K and Simar L, *Applied Multivariate Statistical Analysis*, Springer-Verlag, Berlin Heidelberg, 2012.

[51] Li W and Li J, Modeling and analysis of RNA-seq data: A review from a statistical perspective, *Quantitative Biol.*, 2018, **6**(3): 195–209.

[52] Samuels M L, Witmer J A, and Schaffner A A, *Statistics for the Life Sciences*, 5th Edition, Pearson Education, Edinburgh Gate, Harlow, 2016.

[53] Anders S and Huber W, Differential expression analysis for sequence count data, *Genome Biol.*, 2010, **11**(10): R106.

[54] Love M I, Huber W, and Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.*, 2014, **15**(12): 550.

[55] Li H, Wei Z, and Maris J M, A hidden Markov random field model for genome-wide association studies, *Biostat.*, 2010, **11**: 139–150.

[56] Chen M, Cho J, Zhao H, et al., Incorporating biological pathways via a Markov random field model in genome-wide association studies, *PLoS Genet.*, 2011, **7**: e1001353.

[57] Hou L, Chen M, Zhang C K, et al., Guilt by rewiring: Gene prioritization through network rewiring in genome wide association studies, *Hum. Mol. Genet.*, 2014, **23**(10): 2780–2790.

[58] Chalhoub B, Denoeud F, Liu S, et al., Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome, *Science*, 2014, **345**: 950–953.

[59] Wang X, Wang H, Wang J, et al., The genome of the mesopolyploid crop species Brassica rapa, *Nat Genet.*, **43**: 1035–1039.

[60] Liu S, Liu Y, Yong C, et al., The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes, *Nat. Commun.*, 2014, **5**: 3930.

[61] Huala E, Dickerman A W, Garciahernandez M, et al., The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant, *Nucleic Acids Res.*, 2001, **29**: 102–105.

[62] Li C and Li H, Network-constrained regularization and variable selection for analysis of genomic data, *Bioinformat.*, 2008, **24**(9): 1175–1182.

[63] Liao J G and Chin K V, Logistic regression for disease classification using microarray data: Model selection in a large p and small n case, *Bioinformat.*, 2007, **23**(15): 1945–1951.