



Research Article

Analysis and Research of Key Genes in Gene Expression Network Based on Complex Network

Guobin Chen,¹ Jun Qi,^{2,3} Chao Tang,^{2,3} Ying Wang,^{2,3} Yongzhong Wu ^{2,3},
and Xiaolong Shi ^{2,3}

¹Chongqing Key Laboratory of Spatial Data Mining and Big Data Integration for Ecology and Environment, Rongzhi College of Chongqing Technology and Business University, Chongqing 401320, China

²Radiation & Cancer Biology Laboratory,

Chongqing University Cancer Hospital & Chongqing Cancer Institute & Chongqing Cancer Hospital, Chongqing 400030, China

³Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, Chongqing 400030, China

Correspondence should be addressed to Yongzhong Wu; cqmdwyz@163.com and Xiaolong Shi; xshi.bear@cqu.edu.cn

Received 1 September 2020; Revised 12 October 2020; Accepted 30 October 2020; Published 10 December 2020

Academic Editor: Atif Khan

Copyright © 2020 Guobin Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gene expression network is also a type of complex network. It is challenging to analyze the gene expression network through relevant knowledge and algorithms of a complex network. In this paper, the existing characteristics of genes are analyzed from various indexes of the gene expression network to analyze key genes and TOP genes. Firstly, gene chip data are screened, gene data with obvious characteristics are selected, and relevant clustering characteristics are analyzed. Then, the complex gene network structure is established, and gene networks with different threshold shapes and different sizes are selected. Finally, the relevant indexes and PR values after the PageRank algorithm are analyzed for complex networks under different thresholds, thus establishing the TOP gene and PR sequence.

1. Introduction

With the development of gene chip and second-generation sequencing and the emergence of new technologies, the analysis of the human genetic structure has become a reality, the association information between genes is analyzed, being able to express genes related to key genes, and it is possible to analyze the correlation between genes and diseases more effectively. Establishing a biological network can obviously discover and analyze the correlation and influence between expression elements in various biological systems. It makes it possible to study the characteristics of organisms at a certain level [1]. Unlike the research on inherent single molecules, the information at the whole system level can be displayed by establishing networks.

At present, molecular biology network analysis tools in this field mainly include metabolic network [2], protein-protein

interaction network [3], gene coexpression network [4], gene regulatory network [5], and signal transmission network [6]. In order to expand the scope and depth of the current research, the transition from a single molecule to a network system level becomes the next major development direction of interactive networks because of, in general, the complex biological phenomena and pathology that cannot be caused by only one factor. Therefore, as a tool to explore complex pathological and life phenomena, the interactive network provides an effective analysis method, to perfect the comprehensive research. These networks have great research value, among them; the gene coexpression network has irreplaceable research characteristics in some aspects. Whether the amount of gene expression can establish the correlation of related diseases, the expression network can establish the complex network, and it is challenging to analyze the gene correlation with the related technologies of the complex network.

2. Related Works

The reconstruction algorithm of the gene regulatory network and the exploration of a series of related models began in the middle of the 20th century. Raterproposed the system and characteristics to control the interaction and interaction between all genes in prokaryotic cells exploring the mutual influences and interactions between genes in prokaryotic cells, as well as system relationships and characteristics. In 1969, Kauffman discovered and described the exploration research with epoch-making significance, using popular binary logic rules to stipulate the basic construction model and reconstruction algorithm of gene network [7–9]. D’Haeseleer et al. [10] were the first to use ordinary differential equations to do some research on time series data. Gardner et al. [11] were the first to prove that, by using their multivariate regression network identification (NIR) algorithm, the steady state measurements can be used to infer the network structure. They studied a data set, overexpression of specific genes in bacterial models with plasmids, when the gene expression level reaches a new steady state value, using its measured values. This method has been suggested for transcriptase data sets including siRNA knockout experiments [12]. In the past decades, numerous computational methods have been proposed to infer gene regulatory networks. These methods can be roughly divided into coexpression-based methods [13], supervised learning method [14], the model-based method [15], and the information theory-based method [16, 17] with lower computational complexity. However, direct correlation and model system dynamics cannot be inferred. Based on the supervised learning method, we mainly use the known rules to infer gene regulatory networks on genome-wide data, such as SEREND [14], GENIES [15], and SIRENE [18], but all need additional information of regulatory relationships to cultivate models. By guiding reasoning from prior information of known rules, we can achieve higher accuracy and be superior to many other methods.

Constructing complex gene expression networks and using large-scale gene expression data sets for network analysis are effective methods to reveal new biological knowledge. However, methods for gene association in the construction of these coexpression networks have not been thoroughly evaluated. Because different methods lead to different coexpression networks with different structures and provide different information, it is very important to select the best gene association method. Zhang et al. [19] proposed the identification of protein complexes in protein interaction network (IPC-RPIN), which effectively fused topological structure, gene expression profile, and GO functional annotation information to achieve gene-protein expression complex. Hua et al. [20] proposed the fusion research of three commonly used reasoning algorithms to establish a genome-scale and high-quality gene coexpression network. After applying this expression network to monocotyledonous plant rice, the network quality has been verified and evaluated through the selected gene function association data sets, which is obviously superior to other methods. Li et al. [21] predicted subcellular location information by integrating time-history gene expression data with the spatial and temporal active protein interaction

network (ST-APIN). In order to evaluate the efficiency of the proposed method, the commonly used classical clustering algorithm has obvious advantages in identifying protein complexes in ST-APIN and the other three dynamic PIN.

3. Technical Indicators Related to Complex Networks

3.1. Gene Degree. Degree is an important index to describe the attributes of nodes. The degree of nodes refers to the measurement that genes are associated with genes. In directed complex networks, the degree of nodes can be divided into exit degree and entry degree, which indicates the number of genes pointing to other genes and the number of other genes pointing to the gene. The average degree of a complex network reflects the density program of the network, and the degree distribution can describe the importance of different genes through the number of genes and gene connections. When the scale of gene expression network is very large, the degree distribution of genes can fully display the distribution law of genes and can identify and distribute different types of networks.

The distribution of gene complex network degree is described as follows:

- (A) Gene distribution function $P(x)$: It indicates the proportion of gene x in the gene expression network in the whole expression network.
- (B) Cumulative degree distribution function P_x : The P_x description degree is not less than the probability distribution of gene x , and the distribution function is as follows:

$$P_x = \sum_{k=x}^{\infty} P(k). \quad (1)$$

If the gene degree of the gene expression complex network obeys the distribution function $P(x) \propto k^{-\gamma}$ of the power index γ , it is said that the gene complex network obeys the scale-free network of power-law distribution, while the accumulation degree distribution function P_x obeys the power-law distribution $P_x \propto x^{1-\gamma}$ of power $\gamma - 1$.

3.2. Gene Network Density. Gene network density $d(G)$ is used to describe the density of connection edges between genes in the gene network and is defined as the ratio of the actual number of gene connections in the gene network to the upper limit of the number of gene connections that can be accommodated. Dense programs and dynamic evolution laws used to measure the relationship between gene expression networks in gene complex networks. There are N gene nodes and L edges in the gene complex network, and the density formula (2) of the gene complex network is defined as follows:

$$d(G) = \frac{2L}{N(N-1)}. \quad (2)$$

The density range of the gene complex network is $[0, 1]$. When $d(G) = 1$, the network is all connected. When $d(G) = 0$, there are no connections in the network. In the

actual gene network, the network with density 1 does not exist. In most networks, it is around 0.5. The density of large-scale networks is smaller than that of small-scale networks, and the density of networks of different sizes cannot be directly compared. The absolute density formula (3) can be used to compare the network densities of different sizes:

$$d(G) = \frac{M * 3 D}{4SR^3}. \quad (3)$$

In the above, D represents the diameter of the network, R represents the radius, and S represents the circumference of the network.

3.3. Aggregation Coefficient of Gene Network. Gene network clustering coefficient is used to describe the degree of node neighbors associated with gene nodes in a gene network. In the gene network, the node v_i and the gene network aggregation coefficient C_i represent the probability of interconnection with its adjacent gene nodes. k_i is used to represent the number of neighbors interconnected by gene nodes v_i , and e_i is used to represent the number of interconnected neighbors existing in k_i . $k_i(k_i - 1)/2$ represents the maximum number of interconnections, and equation (4) represents the aggregation coefficient of gene node v_i :

$$C_i = \frac{2e_i}{k_i(k_i - 1)}. \quad (4)$$

The aggregation coefficient of gene nodes is vividly described in the gene complex network, which can be understood as the correlation between one gene node and another gene node, the probability of connection between this gene node and the neighbor of another gene also exists, and the probability comparison of correlation exists, so it has strong aggregation in the gene complex network. The aggregation of the whole gene network is evaluated by the average aggregation coefficient. The average aggregation coefficient of the gene complex network is defined as the average aggregation coefficient of all gene nodes in the gene network. Equation (5) is described as follows:

$$C = \frac{\sum_{i \in v_i} C_i}{V}, \quad (5)$$

where $|V|$ represents the number of gene nodes, C_i represents the aggregation coefficient of gene node v_i , and its value range is $[0, 1]$. When $C = 0$, there are no connected gene nodes in the gene network; when $C = 1$, the gene network node is all connected.

The average aggregation coefficient in the gene expression network describes the probability of connection between any associated genes in the gene network and reflects the closeness of the node relationship of the whole gene.

4. Key Gene Determination Method Based on PageRank Algorithm

4.1. Construction of Complex Network for Gene Expression. Gene chip expression matrix reflects different gene expression levels of different sequencing samples, which can be

described by $n \times m$ matrix, where X_{nm} represents the expression level of the n th gene in m samples:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{bmatrix}. \quad (6)$$

In this paper, the WGCNA algorithm [22] is used to construct the gene expression network. The WGCNA algorithm is based on a scale-free network. Whether there is a correlation between the two genes can be expressed by the correlation coefficient, forming the following formula:

$$M_{ij} = |\text{cor}(X_i, X_j)|, \quad (7)$$

M_{ij} represents the similarity of gene i and gene j in the expression amount of samples in the chip and M represents the similarity matrix in $[0, 1]$. In the construction of the gene expression matrix, the similarity is calculated through the expression of different genes in all samples. The similarity is an important measurement index to measure the strength between genes and is also the basis for the construction of complex networks. In the WGCNA algorithm, S_{ij} is measured by weighting using a soft threshold, and β is used to represent the soft threshold, while the general network structure adopts a hard threshold. The gene expression network using a soft threshold is more prominent, and its strong correlation is more prominent. Equation (8) is as follows:

$$M_{ij} = |\text{cor}(X_i, X_j)|^\beta. \quad (8)$$

The weighting coefficient is defined according to scale network, and the correlation between $\log(k)$ of network node number K and $\log(p(k))$ of node occurrence probability in a scale-free network is required to reach more than 0.8. Considering the correlation between genes and genes, the correlation between the individual genes and genes is expressed by the adjacency matrix, and topological overlap measure (TOM) is expressed by the strong relationship between genes and genes:

$$\text{TOM} = \frac{\sum_{k \neq i, j} M_{ik} M_{kj} + M_{ij}}{\min(\sum_k M_{ik} + \sum_k M_{jk}) + 1 - M_{ij}}. \quad (9)$$

Then, set the correlation threshold λ (generally referring to the hard threshold) and compare it with the value of TOM matrix. If the matrix value in TOM is greater than λ , it indicates that there is a correlation between gene i and gene j (which can be expressed by adjacency matrix $M_{ij} = 1$). If the matrix value in TOM is less than λ , it means that there is no correlation between gene i and gene j (which can be expressed by adjacency matrix $M_{ij} = 0$). The generation of gene coexpression network according to the above principles is also the basis of this study and provides the basic conditions for analyzing core genes.

4.2. Convergence of PageRank Algorithm. In the complex network of genes, there is a correlation between each gene

and the genes related to it, but the influence of genes related to genes should also be considered. Through the correlation between genes, the PageRank algorithm is used to realize this correlation between genes, thus determining key genes.

Feature vector centrality and its variants are widely used. For example, the most famous PageRank algorithm [23] in the field of web page sorting is the core algorithm of Google search engine. At the initial time, each gene is given the same PR value, and then iteration is carried out, and the current PR value of each gene is divided equally to all genes it points to in each step. The new PR value of each gene is the sum of the PR values it obtains, so that the PR value of the node P_i at time t is defined as the following equation:

$$PR_i(t) = \sum_{j=1}^n a_{ij} \frac{PR_j(t-1)}{k_j^{\text{out}}}, \quad (10)$$

where k_j^{out} is the output of node P_j , a_{ij} gene expresses the initial matrix of the correlation network, and PR iterates until the value of each node reaches stability.

The PageRank algorithm is applied in many fields. Through continuous iteration, a relatively stable value is reached. As the number of iterations increases, the error ε of the values generated before and after is expressed. When a reasonable error value ε is reached, the PageRank algorithm ends, as expressed by the following equation:

$$PR_i(t) - PR_i(t-1) < \varepsilon. \quad (11)$$

Theoretically, ε is determined according to the user's experimental results. When the error comparison is required to be small, ε comparison is set. When $i \rightarrow \infty$, $\varepsilon = 0$.

The relevant literature of the PageRank algorithm does not give the relevant convergence proof process, only the iterative process is used to explain its convergence, and no relevant mathematical methods are used to explain it. The following author gives the relevant proof process.

Theorem 1. *If the sum of the values of all columns of a gene correlation matrix is 1, then the matrix is a convergence matrix.*

Proof: Let $PR(t) = [PR_1(t), PR_2(t), \dots, PR_n(t)]'$

$$M = \begin{bmatrix} \frac{a_{11}}{k_1^{\text{out}}} & \frac{a_{12}}{k_2^{\text{out}}} & \dots & \frac{a_{1n}}{k_n^{\text{out}}} \\ \frac{a_{21}}{k_1^{\text{out}}} & \frac{a_{22}}{k_2^{\text{out}}} & \dots & \frac{a_{2n}}{k_n^{\text{out}}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{a_{n1}}{k_1^{\text{out}}} & \frac{a_{n2}}{k_2^{\text{out}}} & \dots & \frac{a_{nm}}{k_n^{\text{out}}} \end{bmatrix}, \quad (12)$$

$$\begin{aligned} PR(t) &= M * PR(t-1), \\ &\Rightarrow PR_i(t) = M^t * PR(1), \end{aligned}$$

in which $\sum_{i=1}^n PR_i(t) = 1$; it is sufficient to prove that M^t in formula (12) converges.

Let $M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ m_{21} & m_{22} & \dots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & \dots & m_{nn} \end{bmatrix}$. The eigenvalue of M is

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}, \quad M = P * \Lambda * P^{-1} \quad (13)$$

$$\Rightarrow M^t = P * \Lambda^t * P^{-1},$$

in which the sum of each column is 1 and $m_{ii} = 0$, which is proved $-1 \leq \lambda_i \leq 1$.

According to the disk theorem, $M = (m_{ij})$ has N disks whose N eigenvalues all fall on the plane:

$$D_i(M) = \{\lambda_i | |\lambda_i - m_{ii}| \leq R_i\}, \quad i = 1, 2, \dots, n, \quad (14)$$

in which $\lambda_i \in \cup_{j=1}^n D_i(M)$, $R_i = \sum_{j=1, j \neq i}^n |m_{ij}|$.

In the matrix M , N disks of M given according to equation (14) are

$$D_i(M) = \{\lambda_i | |\lambda_i| \leq 1\}, \quad i = 1, 2, \dots, n. \quad (15)$$

The N eigenvalues of M^t all fall in the union set of n disks on the complex plane. The disk eigenvalues are seen as shown in Figure 1.

Therefore, when $m = t$ is a convergence matrix, the proof is complete.

λ_i is in the white disk of Figure 1; we can look at N eigenvalues $|\lambda_i| \leq 1$ of M^t , according to equation (13).

$$M^t = P * \begin{bmatrix} \lambda_1^t & 0 & \dots & 0 \\ 0 & \lambda_2^t & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n^t \end{bmatrix} * P^{-1}, \quad \text{where}$$

$$t \rightarrow \infty, \lambda_i^t \in \{-1, 0, 1\}, \text{ so that}$$

$$\lim_{t \rightarrow \infty} M^t = P * \begin{bmatrix} \lim_{t \rightarrow \infty} \lambda_1^t & 0 & \dots & 0 \\ 0 & \lim_{t \rightarrow \infty} \lambda_2^t & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lim_{t \rightarrow \infty} \lambda_n^t \end{bmatrix} * P^{-1},$$

$$\Rightarrow \lim_{t \rightarrow \infty} M^t \Rightarrow P * \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & -1 \end{bmatrix} * P^{-1}. \quad (16)$$

Therefore, when $\lim_{t \rightarrow \infty} M^t$ is a convergence matrix, the proof is complete.

4.3. Disk Analysis of Improved PageRank Algorithm. The above PageRank algorithm is based on the traditional method to prove convergence. Because the traditional β PageRank algorithm does not satisfy the characteristics of strong connectivity and traps, that is, some web pages do not have links and loops pointing to other web pages, there are

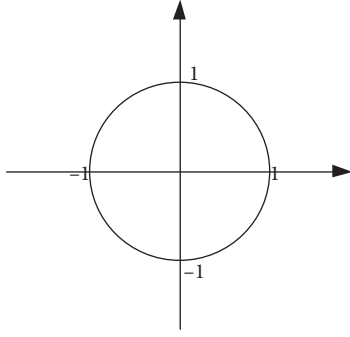


FIGURE 1: Disk eigenvalues.

links pointing to them or links not pointing to other web pages. The improved PageRank algorithm is

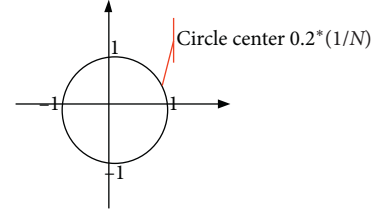


FIGURE 2: PageRank disk.

$$A = \alpha * M + (1 - \alpha) * \left[\frac{1}{N} \right]_{N*N}. \quad (17)$$

M satisfies the normalized matrix of the traditional PageRank algorithm. After β takes 0.8, the form of A matrix is as follows:

$$A = \begin{bmatrix} 0.2 * \frac{1}{N} & 0.8 * m_{12} + 0.2 * \frac{1}{N} & \dots & 0.8 * m_{1n} + 0.2 * \frac{1}{N} \\ 0.8 * m_{21} + 0.2 * \frac{1}{N} & 0.2 * \frac{1}{N} & \dots & 0.8 * m_{2n} + 0.2 * \frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ 0.8 * m_{n1} + 0.2 * \frac{1}{N} & 0.8 * m_{n2} + 0.2 * \frac{1}{N} & \dots & 0.2 * \frac{1}{N} \end{bmatrix}. \quad (18)$$

At this time, the convergence of matrix A is proved with $0.2 * (1/N)$ as the center of the circle, as shown in Figure 2.

As can be seen from Figure 2, the center of the improved PageRank algorithm is $0.2 * (1/N)$, and the radius is $0.8 + 0.2 * (N - 1/N)$ or $1 - 0.2 * (1/N)$, so that the disk can pass through the point of 1, but the disk does not pass through the point of -1, so the maximum eigenvalue of the improved PageRank transfer matrix takes 1 and the minimum value cannot take -1. Therefore, when $t \rightarrow \infty$, $\lambda_i^t \in \{0, 1\}$ is easier to converge.

4.4. Calculation of PR Value When $t \rightarrow \infty$. Through the above proof, the PageRank algorithm is convergent. How to find the PR value is a problem studied by many scholars, especially in the improved PageRank algorithm, there are a large amount of literature, and some scholars have done relevant work on how to quickly find the PR value and accelerate convergence. This section proposes a PR value calculation method when $t \rightarrow \infty$. When $t \rightarrow \infty$, the PR value is the actual value. Iterative calculation of the PR value through other PageRank algorithms is an approximate value. The accuracy cannot be compared with the PR value when $t \rightarrow \infty$. The following calculation process is given. The formula of the PageRank algorithm is as follows:

$$R(N) = M * R(N - 1). \quad (19)$$

Let $R(N) - R(N - 1) \leq \varepsilon$, when $\lim_{N \rightarrow \infty} \varepsilon = 0$, $R(N) = R(N - 1)$, the formula became

$$[R_1, R_2, R_3, \dots, R_n]' = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} * [R_1, R_2, R_3, \dots, R_n]'. \quad (20)$$

Namely,

$$\begin{cases} 0 + a_{12} * R_2 + a_{13} * R_3 + \dots + a_{1n} * R_n = R_1 (1), \\ a_{21} * R_1 + 0 + a_{23} * R_3 + \dots + a_{2n} * R_n = R_2 (2), \\ \vdots \\ a_{n1} * R_1 + a_{n2} * R_2 + a_{n3} * R_3 + \dots + 0 = R_n (n). \end{cases} \quad (21)$$

Theorem 2. When $t \rightarrow \infty$, the polynomials (1), (2), (3), ..., (n) are linearly correlated.

Proof. The sum of the former $n-1$, $(1) + (2) + \dots + (n-1) = (a_{21} + a_{31} + \dots + a_{n-1,1}) * R_1 + (a_{12} + a_{32} + \dots + a_{n-1,2}) * R_2 + (a_{13} + a_{23} + \dots + a_{n-1,3}) * R_3 + \dots + (a_{1n} + a_{2n} + \dots + a_{n-1,n}) * R_n = R_1 + R_2 + \dots + R_{n-1}$.

When $\sum_{j=1}^n a_{ji} = 1$,

$$\begin{aligned}
&\Rightarrow (1 - a_{n1}) * R_1 + (1 - a_{n2}) * R_2 + (1 - a_{n3}) * R_3 + \dots \\
&\quad + (1 - 0) * R_n = R_1 + R_2 + \dots + R_{n-1} \\
&\Rightarrow R_1 + R_2 \dots + R_n - a_{n1} * R_1 - a_{n2} * R_2 - a_{n3} \\
&\quad * R_3 \dots - a_{nm-1} * R_{n-1} = R_1 + R_2 + \dots + R_{n-1} \\
&\Rightarrow a_{n1} * R_1 + a_{n2} * R_2 + a_{n3} * R_3 + \dots + a_{m-1} \\
&\quad * R_{n-1} = R_n(n).
\end{aligned} \tag{22}$$

From the above, it can be seen that polynomial (1) + (2) + (3) \dots + (n - 1) = (n) holds, and it is deduced that any $N-1$ correlation is equal to another term, so the above polynomial is linearly correlated. \square

Theorem 3. When $t \rightarrow \infty$, polynomials (1), (2), (3), \dots , (n - 1) are linearly independent.

Proof. Assume that k_1, k_2, \dots, k_{n-2} makes the first $n - 2$ term represent the $n - 1$ term; that is,

$$k_1 * (1) + k_2 * (2) \dots + k_{n-2} * (n - 2) = (n - 1). \tag{23}$$

According to Theorem 1, it is concluded that (1) + (2) \dots + (n - 2) + (n) = (n - 1) is equal on both sides:

$$\begin{aligned}
&k_1 * (1) + k_2 * (2) \dots + k_{n-2} * (n - 2) \\
&= (1) + (2) \dots + (n - 2) + (n) \\
&\Rightarrow k_1 - 1 * (1) + (k_2 - 1) * (2) \dots + (k_{n-2} - 1) * (n - 2) \\
&= (n).
\end{aligned} \tag{24}$$

According to Theorem 1, it is concluded that (1) + (2) \dots + (n - 2) + (n - 1) = (n) is equal on both sides:

$$\begin{aligned}
&\Rightarrow k_1 - 1 * (1) + (k_2 - 1) * (2) \dots + (k_{n-2} - 1) * (n - 2) \\
&= (1) + (2) \dots + (n - 2) + (n - 1) \\
&\Rightarrow k_1 - 2 * (1) + (k_2 - 2) * (2) \dots + (k_{n-2} - 2) * (n - 2) \\
&= (n - 1),
\end{aligned} \tag{25}$$

n equal to equation (11).

$$\Rightarrow (1) + (2) \dots + (n - 2) + (n - 2) = 0. \tag{26}$$

According to Theorem 1, the conclusion is (1) + (2) \dots + (n - 2) + (n - 1) = (n) = > (n - 1) = (n).

Similarly, it is possible to obtain that all multiple terms are equal, which is contradictory to the problem, so (1), (2), (3), \dots , (n - 1) are linearly independent.

According to the properties of the PageRank algorithm, $R_1 + R_2 + R_3 + \dots + R_n = 1$ takes any polynomial $n-1$ term and combines it with the above:

$$\begin{cases} -R_1 + a_{12} * R_2 + a_{13} * R_3 + \dots + a_{1n} * R_n = 0(1) \\ a_{21} * R_1 - R_2 + a_{23} * R_3 + \dots + a_{2n} * R_n = 0(2) \\ \vdots \\ R_1 + R_2 + R_3 + \dots + R_n = 1(n). \end{cases} \tag{27}$$

After (n) terms are added, the above N terms are linearly independent. Equation (27) can be transformed into

$$A' * R = \begin{bmatrix} -1 & a_{12} & \dots & a_{1n} \\ a_{21} & -1 & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} * [R_1 R_2 R_3, \dots, R_n]' \tag{28}$$

$$= [000 \dots 1]'$$

A' matrix is reversible, and the value of R is

$$[R_1 R_2 R_3, \dots, R_n]' = A'^{-1} * [000 \dots 1]'. \tag{29} \quad \square$$

5. Experimental Analysis

5.1. Data Selection. The SVM-REF algorithm was used to select 2105 genes and 265 samples (198 TNBC, 67 non-TNBC). By establishing decision trees, we can see the distribution of decision trees under 2105 genes for each sample, as shown in Figure 3.

As can be seen from Figure 3, GSM1974605 has outliers under 2105 genes, and the GSM1974605 sample can be deleted. After deleting outlier samples, it can be seen that they are roughly divided into two categories, and the distinction is obvious, indicating that 2105 gene selection is reasonably classified under 197 samples.

Selecting soft threshold B plays a very important role in forming TOM gene association matrix. In this paper, the optimal threshold is obtained through continuous iteration of the threshold. Since the scale-free network atlas structure can reach 0.8 or the average connectivity can reach below 100 when $\beta < 15$, if the above requirements cannot be met, the experimental results will be greatly affected. Figure 4 is the distribution diagram of the threshold value and average connectivity.

As can be seen from Figure 4, at that time, the network map structure could reach more than 0.9, indicating that the threshold value of this interval is ideal and the optimal value is the most 8. However, the average connectivity is less than 100 at 3, so $\beta = 8$ meets the requirements of the above conditions.

The TOM matrix is established. The TOM matrix is an index value representing the correlation between genes and is converted into a connection matrix through the TOM matrix. The relevant thresholds of TOM matrix are screened and the thresholds are introduced to form different complex gene networks after continuous changes. As that threshold value = 0.2186, a complex network of gene is shown in Figure 5; as that threshold value = 0.2664, a complex network of gene is shown in Figure 6.

As can be seen from Figures 5 and 6, when the threshold B increases, the complex network diagram gradually decreases, the network structure becomes smaller, and the network community will appear. Smaller outlier gene

networks will also appear so as B increases to a certain value, the network structure changes from a large-scale complex network to a small-scale network structure.

5.2. Analysis of Complex Network Structure. The analysis of complex network structure analyzes the structural changes of the gene complex network from the indexes of centrality, aggregation coefficient, central potential, and so forth. By introducing threshold λ to change the structure of the gene complex network, the network structure can be seen to change continuously by increasing the threshold λ , and the network changes from a large-scale structural model to a small-scale structural model. The specific effect is shown in Figures 7–10.

From Figures 7–10, it can be seen that, with the increase of λ , the structure of the gene network has changed greatly. With the increase of the performance of the network structure, the scale of the whole network is continuously decreasing, resulting in the continuous changes of various complex network parameters, the overall scale is small, and the network community is increasing, resulting in the continuous changes of key gene nodes. Then, PageRank is used to calculate network structures of different sizes, and PageRank values of genes are sorted. Under different threshold settings, the PageRank values of each gene node are shown in Tables 1–4.

Tables 1–4 show the PR values of genes in the gene complex network under different λ values. With the continuous increase of λ , the number of genes in the gene complex network is continuously decreasing, and the expressed gene network structure is also in different states. The values of gene PR values in different networks are also constantly changing, and the ranking of gene PR values is also constantly changing. In each threshold, it is the calculation of different screening networks, which can be determined according to the size of the network or the number of genes. When the number of selected genes is large, you can choose $\lambda = 0.2345$ or $\lambda = 0.2505$. If λ is relatively small, the number of selected genes will be relatively large, and there will be more key genes. When λ is larger, the relative number is small, but the key genes are few. Therefore, when choosing, you can choose several suitable values of λ for comparison. For example, a certain gene appears in several tables consecutively.

In Table 1 to Table 4, several TOP gene tables are ranked by PR value, and the top gene is the key gene. The PR values are given in the table, and the sum of the PR values of all genes is 1.

Figure 11 shows the changes of several gene values (CDK13, DSPP, HLA.G, LINC00304, TPX2, FOXM1) and screens several genes with a higher ranking.

As can be seen from Figure 11, as the λ value and PR value change continuously, the whole PR value increases continuously. As the network scale becomes smaller, the PR value of gene nodes increases continuously. Some genes become outliers because the network scale is constantly changing, and gene nodes will not have new PR values in the later screening process, resulting in PR values of 0.

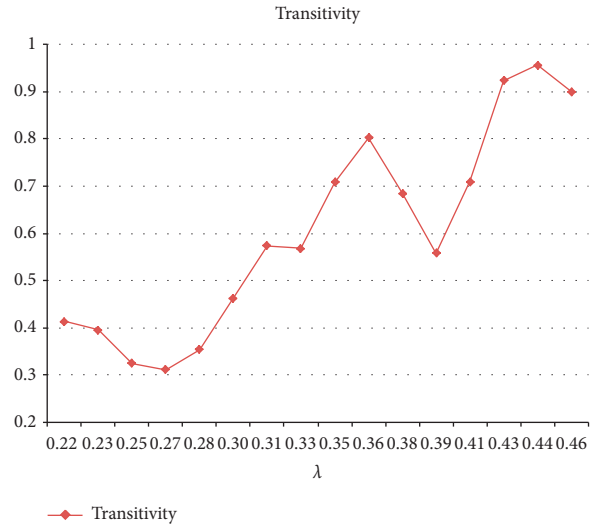


FIGURE 7: Effect diagram of transitivity change after threshold λ is increased.

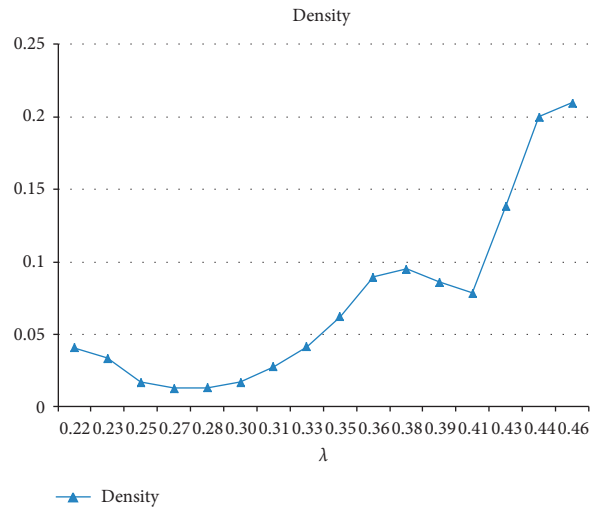


FIGURE 8: Effect diagram of density change after threshold λ is increased.

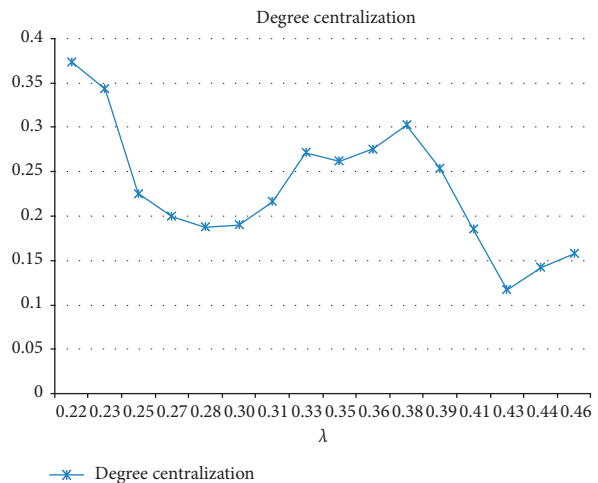


FIGURE 9: Effect diagram of degree centralization after threshold λ is increased.

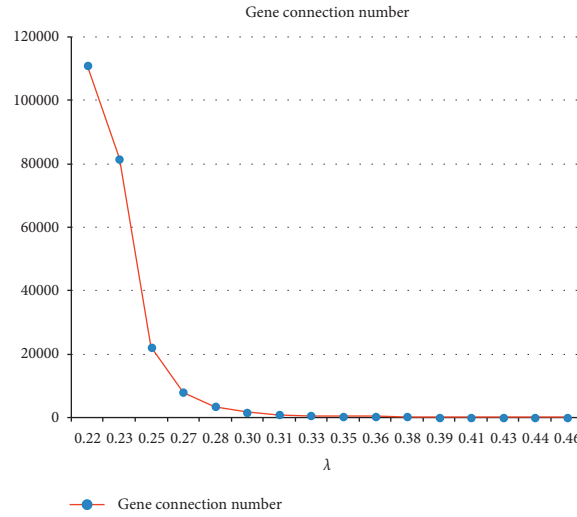
FIGURE 10: Effect diagram of gene connection number change after threshold λ is increased.

TABLE 1: TOP gene and PR values with different thresholds under the PageRank algorithm.

$\lambda = 0.2186$ (TOP10)		$\lambda = 0.2345$ (TOP10)		$\lambda = 0.2505$ (TOP10)		$\lambda = 0.2664$ (TOP10)	
KeyGene	PR value	KeyGene	PR value	KeyGene	PR value	KeyGene	PR value
TBC1D9	0.0056466	TBC1D9	0.006586	TBC1D9	0.011851	BAG6	0.019218
CDC20	0.0050929	HDGF	0.005776	CDC20	0.010072	TBC1D9	0.015549
HDGF	0.0050338	CDC20	0.00561	BAG6	0.010048	CDC20	0.015447
INPP4B	0.0046938	INPP4B	0.005545	HDGF	0.00879	VARs	0.012313
MLPH	0.0043071	MLPH	0.00498	MLPH	0.008688	GNB2	0.012045
KLHDC1	0.0040236	FOXMI	0.00452	INPP4B	0.00861	FOXMI	0.011555
FOXMI	0.0039189	BAG6	0.004413	VARs	0.007983	MLPH	0.011512
PIMREG	0.0038781	KLHDC1	0.004392	FOXMI	0.007928	TPX2	0.009818
BAG6	0.0037029	VARs	0.004176	TPX2	0.006754	AGR3	0.008992
AGR3	0.0036564	AGR3	0.004174	PIMREG	0.006411	HGS	0.00875

TABLE 2: TOP gene and PR values with different thresholds under the PageRank algorithm.

$\lambda = 0.2823$ (TOP10)		$\lambda = 0.2983$ (TOP10)		$\lambda = 0.3142$ (TOP10)		$\lambda = 0.3301$ (TOP10)	
KeyGene	PR value	KeyGene	PR value	KeyGene	PR value	KeyGene	PR value
BAG6	0.025208	BAG6	0.0234342	FOXMI	0.022818	FOXMI	0.030767
GNB2	0.022289	GNB2	0.0219569	CDC20	0.018771	MLPH	0.02873
TBC1D9	0.017161	TBC1D9	0.0190973	AGR3	0.017939	TPX2	0.028
CDC20	0.017141	MLPH	0.0185887	TPX2	0.017931	CDC20	0.025255
MLPH	0.014081	FOXMI	0.0182001	GNB2	0.015257	AGR3	0.016934
FOXMI	0.014026	CDC20	0.015556	BAG6	0.014758	UBE2C	0.016034
VARs	0.01302	TPX2	0.0153193	MLPH	0.013349	RPL36	0.015991
AGR3	0.012785	HGS	0.0137005	TBC1D9	0.012691	BAG6	0.015663
HGS	0.012107	AGR3	0.0134952	KIF23	0.011256	KIF23	0.015176
TPX2	0.010469	ESR1	0.0094534	HGS	0.010413	CCNA2	0.014671

Therefore, different gene complex network structure models will produce different gene structures, and selecting an appropriate gene structure network will obtain different key genes.

The key genes can be analyzed from the above, and the expression significance of the key genes screened will be explained by biological significance. The expression significance of some key genes is shown in Table 5.

TABLE 3: TOP gene and PR values with different thresholds under the PageRank algorithm.

$\lambda = 0.3461$ (TOP10)		$\lambda = 0.3620$ (TOP10)		$\lambda = 0.3779$ (TOP10)		$\lambda = 0.3939$ (TOP10)	
KeyGene	PR value	KeyGene	PR value	KeyGene	PR value	KeyGene	PR value
MLPH	0.030533	TPX2	0.036319	TPX2	0.051559	TPX2	0.081729
FOXM1	0.030396	MLPH	0.030315	ZNF721	0.031488	KIF20A	0.04712
TPX2	0.028037	FOXM1	0.027117	CDC20	0.029894	FOXM1	0.039724
CDC20	0.024861	ZNF721	0.026409	FOXM1	0.029894	CDCA5	0.034104
ZNF721	0.019968	CDC20	0.024967	KIF20A	0.029472	NUSAP1	0.03389
KIF23	0.018408	KLK7	0.02366	CDCA5	0.027509	CDK13	0.030679
KLK7	0.01789	RPL36	0.02354	CCNA2	0.024692	ZNF721	0.030679
RPL36	0.017798	KIF20A	0.021769	CDK13	0.024609	HLA.G	0.025
PHB2	0.017798	KIF23	0.021411	UBE2C	0.023545	HLA.B	0.025
PUF60	0.017798	CCNA2	0.021411	NCAPH	0.021691	HLA.J	0.025

TABLE 4: TOP gene and PR values with different thresholds under the PageRank algorithm.

$\lambda = 0.4098$ (TOP10)		$\lambda = 0.4257$ (TOP10)		$\lambda = 0.4417$ (TOP10)		$\lambda = 0.4576$ (TOP10)	
KeyGene	PR value	KeyGene	PR value	KeyGene	PR value	KeyGene	PR value
TPX2	0.131757	ZNF721	0.060074	CDK13	0.088563	CDK13	0.097796
ZNF721	0.039423	CDK13	0.057162	ZNF721	0.066437	HLA.G	0.070896
CDK13	0.037512	C9orf64	0.057162	HLA.G	0.0625	HLA.B	0.070896
C9orf64	0.037512	SCARF2	0.056231	HLA.J	0.0625	HLA.C	0.070896
HLA.G	0.03125	LINC00304	0.056231	HLA.E	0.0625	HLA.A	0.070896
HLA.J	0.03125	HLA.G	0.047619	HLA.B	0.0625	DSPP	0.066667
HLA.E	0.03125	HLA.J	0.047619	DSPP	0.0625	LINC00304	0.066667
HLA.B	0.03125	HLA.E	0.047619	LINC00304	0.0625	SCARF2	0.066667
KLK7	0.03125	HLA.A	0.047619	SCARF2	0.0625	MT1G	0.066667
KLK8	0.03125	HLA.C	0.047619	MT1G	0.0625	MT1HL1	0.066667

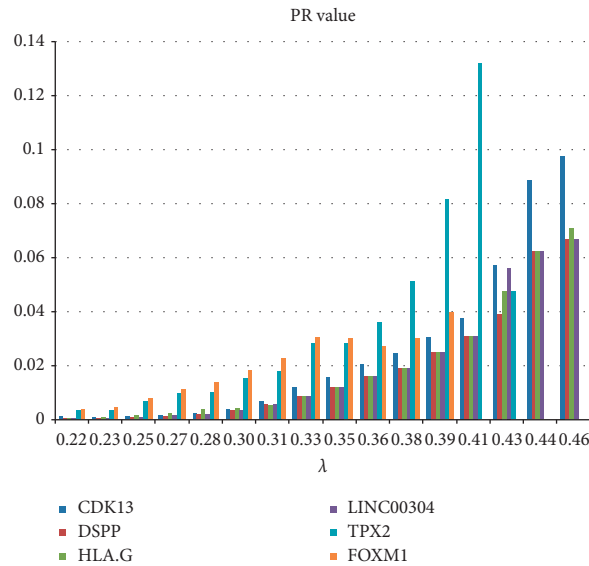


FIGURE 11: Comparison of PR values of genes with different values.

TABLE 5: The biological significance of some key genes.

Gene	The significance of key genes
CDK13	Cyclin-dependent kinase (CDK) 12 and 13 are highly selective dual inhibitors that can inhibit triple-negative breast cancer (TNBC) cells. In terms of mechanism, the inhibition or loss of CDK12/CDK13 will cause the polyadenylation sites of introns to break, thereby inhibiting the expression of critical DNA damage response proteins. This causes the “BRCAness” phenotype, which leads to defects in DNA damage repair and promotes synergy with DNA-damaging chemotherapy and PARP inhibitors.
TPX2	TPX2 protein is highly expressed in breast cancer and has a certain correlation with the histological grade of breast cancer and lymph node metastasis. It may be a risk factor for the occurrence, development, and lymph node metastasis of breast cancer.
FOXM1	FOXM1 is considered as an emerging molecular target with its carcinogenic effect and 85% high overexpression characteristics in TNBC. FOXM1 occupies an important position in the gene network and is a key regulatory gene for breast cancer.

There are many key genes that can be regulated, and only some are listed in Table 5. Other regulated genes can be determined according to the ranking of PR values in Table 1 to Table 4.

6. Conclusion

In this paper, by establishing a complex gene network, statistical analysis is carried out through gene nodes, and then PR values are calculated for each gene node and ranked statistically to obtain gene key nodes. Under different thresholds, the PR value changes of different gene nodes are taken, and the appropriate gene network is selected for screening. The method proposed in this paper can transform the expression amount of gene expression network into a complex gene network and then into PR value, thus obtaining key genes. The next step of research work is to identify key genes from RNA-Seq analysis of expression in the second-generation sequencing and to identify key genes in combination with SNP, InDel, and other variants. It can also consider using other methods [24, 25] to solve the key genes.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Authors' Contributions

Guobin Chen and Jun Qi contributed equally to this work.

Acknowledgments

This work was supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant no. KJZD-K201902101) and the Open Fund of Chongqing Key Laboratory of Spatial Data Mining, and Big Data Integration for Ecology and Environment, Humanities and Social Sciences Project of Rongzhi College of Chongqing Technology and Business University (Grant no. 20197004). The work was also supported by the project (no. cstc2018jxjl10001) from the Natural Science Foundation of Chongqing, the Project Platform Enhancement of Radiation and Cancer Biology Laboratory from Special Funds for Guiding Local Scientific and Technological Development by the Central Government of China, the Project Integrated Innovation and Application of Key Technologies for Precise Prevention and Treatment of Primary Lung Cancer (no. 2019ZX002) from Chongqing Municipal Health Committee, and the Project Technology Platform Construction of Next Generation Sequencing and Research on Clinical Translation from Chongqing Cancer Institute. The funders only provided financial support and did not have any additional role in the study design, data

collection and analysis, decision to publish, or manuscript preparation.

References

- [1] S. P. Ficklin and F. A. Feltus, "Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice," *Plant Physiology*, vol. 156, no. 3, pp. 1244–1256, 2011.
- [2] C. J. Norsigian, K. Erol, S. Yara et al., "iCN718, an updated and improved genome-scale metabolic network reconstruction of acinetobacter baumannii AYE," *Frontiers in Genetics*, vol. 9, p. 121, 2018.
- [3] N. T. Suresh, E. R. Vimina, and U. Krishnakumar, "Multi-scale top-down approach for modelling epileptic protein-protein interaction network analysis to identify driver nodes and pathways," *Computational Biology and Chemistry*, vol. 88, Article ID 107323, 2020.
- [4] J. Li, H. Y. Wang, and C. Xu, "Gene selection of rat hepatocyte proliferation using adaptive sparse group lasso with weighted gene co-expression network analysis," *Computational Biology and Chemistry*, vol. 80, pp. 364–373, 2019.
- [5] B. Van de Sande, C. Flerin, K. Davie et al., "A scalable SCENIC workflow for single-cell gene regulatory network analysis," *Nature Protocols*, vol. 15, no. 7, pp. 2247–2276, 2020.
- [6] M. E. Sommer, J. Selent, C. J. De Graaf et al., "The European research network on signal transduction (ERNEST): toward a multidimensional holistic understanding of G protein-coupled receptor signaling," *ACS Pharmacology & Translational Science*, vol. 3, no. 2, pp. 361–370, 2020.
- [7] S. A. Gloriam, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.
- [8] S. Yelovitch, J. Camden, G. A. Weisman, and B. Fischer, "Boranophosphate isoster controls P2Y-receptor subtype selectivity and metabolic stability of dinucleoside polyphosphate analogues," *Journal of Medicinal Chemistry*, vol. 55, no. 1, pp. 437–448, 2012.
- [9] G. La Regina, R. Bai, W. Rensen et al., "Design and synthesis of 2-Heterocyclyl-3-arylthio-1H-indoles as potent tubulin polymerization and cell growth inhibitors with improved metabolic stability," *Journal of Medicinal Chemistry*, vol. 54, no. 24, pp. 8394–8406, 2011.
- [10] P. D'Haeseleer, X. Wen, S. Fuhrman et al., "Linear modeling of m RNA expression levels during CNS development and injury," *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, vol. 4, p. 41, 1999.
- [11] T. S. Gardner, B. D. Di, D. Lorenz et al., "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, no. 5629, p. 102, 2003.
- [12] D. Hurlley, H. Araki, Y. Tamada et al., "Gene network inference and visualization tools for biologists: application to new human transcriptome datasets," *Nucleic Acids Research*, vol. 40, no. 6, pp. 2377–2398, 2012.
- [13] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.
- [14] H. K. Yalamanchili, B. Yan, M. J. Li et al., "DDGni: dynamic delay gene-network inference from high-temporal data using gapped local alignment," *Bioinformatics*, vol. 30, no. 3, p. 377, 2014.
- [15] M. Kotera, Y. Yamanishi, Y. Moriya, M. Kanehisa, and S. Goto, "GENIES: gene network inference engine based on

- supervised analysis,” *Nucleic Acids Research*, vol. 40, no. W1, pp. W162–W167, 2012.
- [16] J. Zola, A. M. Aluru, and S. Aluru, “Parallel information-theory-based construction of genome-wide gene regulatory networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 12, pp. 1721–1733, 2010.
- [17] N. J. Hudson, N. S. Marina, P. N. Laercio et al., “A haplotype information theory method reveals genes of evolutionary interest in European vs. Asian pigs,” *Journal of Animal Science*, vol. 96, 2018.
- [18] D. Marbach, J. C. Costello, J. C. Costello et al., “Wisdom of crowds for robust gene network inference,” *Nature Methods*, vol. 9, no. 8, p. 796, 2012.
- [19] W. Zang, J. Xu, Y. Li et al., “Integrating network topology, gene expression data and GO annotation information for protein complex prediction,” *Journal of Bioinformatics and Computational Biology*, vol. 17, no. 1, Article ID 1950001, 2019.
- [20] Y. Hua, J. Bingke, L. Lu et al., “NetMiner-an ensemble pipeline for building genome-wide and high-quality gene co-expression network using massive-scale RNA-seq samples,” *Plos One*, vol. 13, no. 2, Article ID e0192613, 2018.
- [21] M. Li, R. X. Meng, Y. F.-X. LiWu, and J. Wang, “Identification of protein complexes by using a spatial and temporal active protein interaction network,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 3, pp. 817–827, 2020.
- [22] P. Pan and S. Horvath, “WGCNA: an R package for weighted correlation network analysis,” *Bmc Bioinformatics*, vol. 9, no. 1, p. 559, 2008.
- [23] I. Rogers, “The google pagerank algorithm and how it works,” *IEEE Communications Letters*, vol. 2, no. 2, pp. 36–38, 2005.
- [24] W. Liu, J. Li, L. Ren, J. Xu, C. Li, and S. Li, “Exploring livelihood resilience and its impact on livelihood strategy in rural China,” *Social Indicators Research*, vol. 150, no. 3, pp. 977–998, 2020.
- [25] W. Jiang, D. R. Carter, H. L. Fu et al., “The impact of the biomass crop assistance program on the United States forest products market: an application of the global forest products model,” *Forests*, vol. 10, no. 3, pp. 1–12, 2019.