Hindawi
BioMed Research International
Volume 2018, Article ID 5670210, 12 pages
https://doi.org/10.1155/2018/5670210

Research Article

Pathway Network Analysis of Complex Diseases Based on Multiple Biological Networks

Fang Zheng , Le Wei, Liang Zhao, and FuChuan Ni

College of Informatics, Huazhong Agricultural University, Wuhan 430079, China

Correspondence should be addressed to FuChuan Ni; fcni_cn@mail.hzau.edu.cn

Received 10 November 2017; Revised 6 February 2018; Accepted 11 March 2018; Published 30 July 2018

Academic Editor: Tao Huang

Biological pathways play important roles in the development of complex diseases, such as cancers, which are multifactorial complex diseases that are usually caused by multiple disorders gene mutations or pathway. It has become one of the most important issues to analyze pathways combining multiple types of high-throughput data, such as genomics and proteomics, to understand the mechanisms of complex diseases. In this paper, we propose a method for constructing the pathway network of gene phenotype and find out disease pathogenesis pathways through the analysis of the constructed network. The specific process of constructing the network includes, firstly, similarity calculation between genes expressing data combined with phenotypic mutual information and GO ontology information, secondly, calculating the correlation between pathways based on the similarity between differential genes and constructing the pathway network, and, finally, mining critical pathways to identify diseases. Experimental results on Breast Cancer Dataset using this method show that our method is better. In addition, testing on an alternative dataset proved that the key pathways we found were more accurate and reliable as biological markers of disease. These results show that our proposed method is effective.

1. Introduction

The growth in knowledge of large-scale transcriptomic and proteomic technologies has enabled the identification of risk factors of complex diseases, personalized medicine, and so forth. Many algorithms (such as the supervision, nonsupervision, and statistics method) have been developed to process these data for acquiring important biological biomarkers. However, these methods still have some limitations and challenges. First, transcriptomic data analysis is the inherent complexity of multiple biological processes. Second, the data from different platforms also lead to noise. Although some methods were used to reduce the deviation, it is difficult to obtain robust result of these data. To circumvent these limitations, some computational methods project gene expression data into a molecular signaling network, but errors generated by variations in experiment also affect the accuracy to distinct different samples.

Biological networks are powerful resources for the discovery of genes and genetic modules that drive disease. Fundamental to network analysis is the concept that genes underlying the same phenotype tend to interact; this principle can be used to combine and to amplify signals from individual genes [1], a biological pathway which plays an important role in understanding the mechanisms of complex diseases, improving clinical treatment, and discovering drug targets and biomarkers [2]. The increasing availability of high-throughput biological data of complex diseases and the development of various biological networks provided better conditions to build accurate pathway analysis models. But due to a lack of abundant pathway knowledge, most pathway analysis results are incomplete, unreliable, or inaccurate [3]. So, it is an important work to build pathway analysis method of multitype data, such as gene expression data, transcriptomic data, and protein data. The major advantage of pathway-based methods is their capability to perform biologically relevant dimension reduction as a result of the analysis [4].

There are some popular pathway analysis tools, such as GSEA, SPIA, DAVID, and Pathologist. They provide different methods to explain the function of the pathway analysis.

GSEA (Gene Set Enrichment Analysis) [5] processes expression profile data with labels, sorting every pathway according to enrichment statistics of the difference of gene.

SPIA (Significant Pathway Inference Analysis) [6] combines the difference expression and pathway structure information. The effect of alteration of gene expression at different positions in the pathway is considered to be different.

DAVID [7] (Database for Annotation, Visualization and Integrated Discovery) can provide a comprehensive functional annotation of the gene list to help researchers understand the biological significance of genes behind. At present, it is the most widely used method of gene function annotation.

Although many of the above methods have shown encouraging results for finding new information, there are still some limitations. In those methods, pathway is simplified into a simple set of genes, treating pathways as unstructured sets of genes, ignoring the functional relationship between different genes in the pathway, so it cannot accurately assess function change of the related pathways.

To overcome the above limitations, new analytical methodologies are required that infer complex transcriptomic changes more accurately into the biologically network.

System biology considers that biological functions are not the result of a single gene or protein, but the interaction result of multiple biological molecules with each other. With the development of system biology, the biological network has become a powerful tool to research the complex biological activity. Biological networks can simultaneously study the interaction relationships between different biological molecules. System biology can help us understand the exercise process of biological function and explore the underlying mechanisms of biological processes.

Because changes in biological function are the result of molecular interactions, the functional annotation of differentially expressed genes should consider not only the effects of differentially expressed genes on the pathway, but also the effects of gene interaction on different pathways. In addition, consideration should also be given to the association between pathways.

So, in this paper, we proposed a network-based pathway analysis method. We find the disease related pathways through the analysis of the constructed network. The specific process for the construction of networks includes the following. First, we integrated protein-protein interaction (PPI) information and gene expression profile data into the pathway, and then the candidate genes associated with disease phenotypes were screened using mutual information calculations. Secondly, we integrated gene's GO information into pathway to calculate the correlation between the pathways and then construct the pathway network. Finally, the critical pathways are identified in the network. The experimental results of this method with breast cancer data show that our method can not only find the high risk of gene and signaling pathways, but also find an association between the risk pathways.

2. Methods

Figure 1 describes our network-based pathway analysis method. First, the differentially expressed genes were identified by comparing the disease samples and normal samples, and then they were projected onto protein interaction data. If two protein nodes all appear in differentially expressed genes, they are preserved, and the candidate genes were screened out. Secondly, the biological signaling pathways are obtained from database MSigDB which includes 1329 sets of biological metabolism and signaling pathways. The candidate genes are projected into the pathway, and we calculate the active score of each pathway for every sample according to the document [8]. Then according to the activity vectors of each biological pathway, combining the phenotypic information of the samples, the mutual information between the activity vectors and the phenotypic vectors of the samples is calculated. Next, the semantic similarity of differentially expressed genes in each pathway is calculated using gene's GO knowledge, and we calculate the correlation between the pathways and then construct the pathway network. At last, we used GeneRank algorithm to find the critical pathways in the network. In the following, the detailed explanations of our proposed method are described.

2.1. Project the Candidate Gene to the Pathway. Gene expression data often aim to identify genes that are differentially regulated across different classes of samples, for example, finding the genes affected by a treatment, or finding marker genes that discriminate diseased from healthy subjects. Using gene expression profiles obtained from a number of genes for several samples or experimental conditions, we can obtain a gene set that shows a differential expression pattern across different samples. However, a differential expression gene set does not guarantee the existence of a real interaction between the corresponding proteins. Instead, it only suggests that there may be an interaction between the proteins. To accurately describe the change in gene interactions for several samples or experimental conditions, here we screen genes with PPI network.

In the PPI network, if the two nodes (gene) of the edge are both in the set of differentially expressed genes, then the two genes were reserved; otherwise they were removed. So we get the candidate gene set (CG set). The biological pathway is a complete metabolic pathway which contains all genes that constitute a set. The gene-gene interaction of pathway is different in different tissues or samples. These differences may be caused by genes interactions of the pathway or neighbor pathway. To find the related biological pathway of the disease, we analyze the expression of each gene set in different sample, calculating the activity score of the biological pathway (active fraction) according to [8]; main procedures are as follows and shown in Figure 2.

(1) We get the PPI data from HPRD (human protein reference database) and gene expression data of the breast cancer from NCBI GEO (the Gene Expression Omnibus at the National Center for Biotechnology Information (NCBI)).

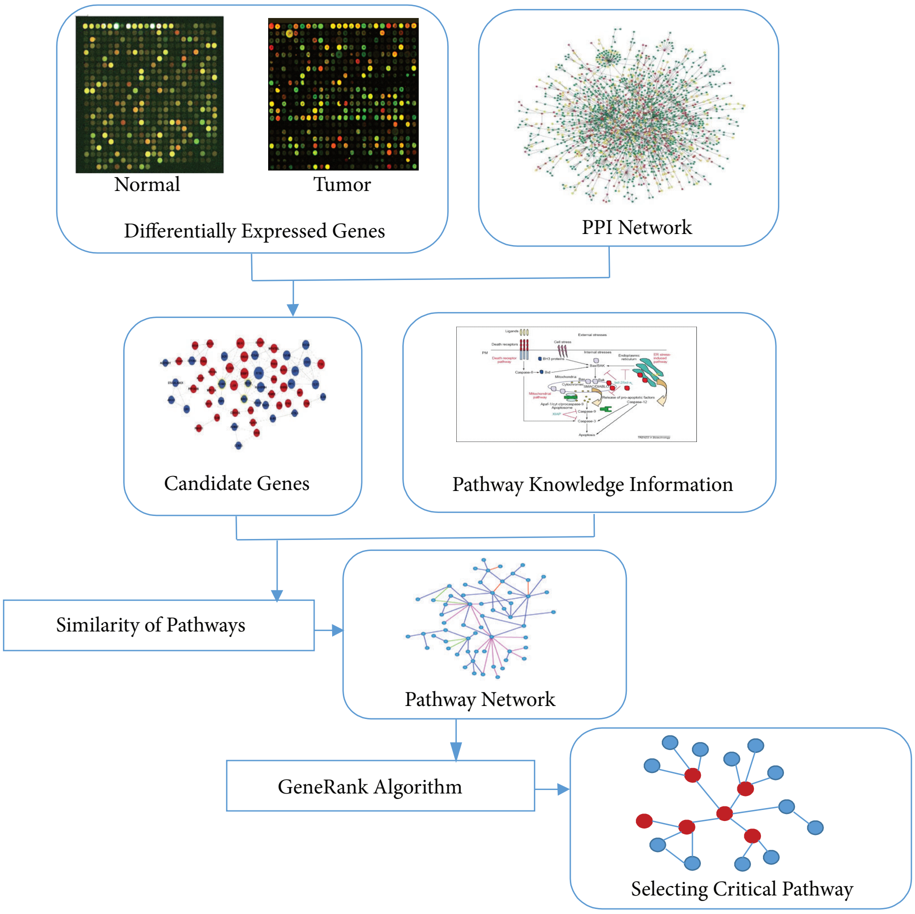

Figure 1: Flow chart of analysis method.

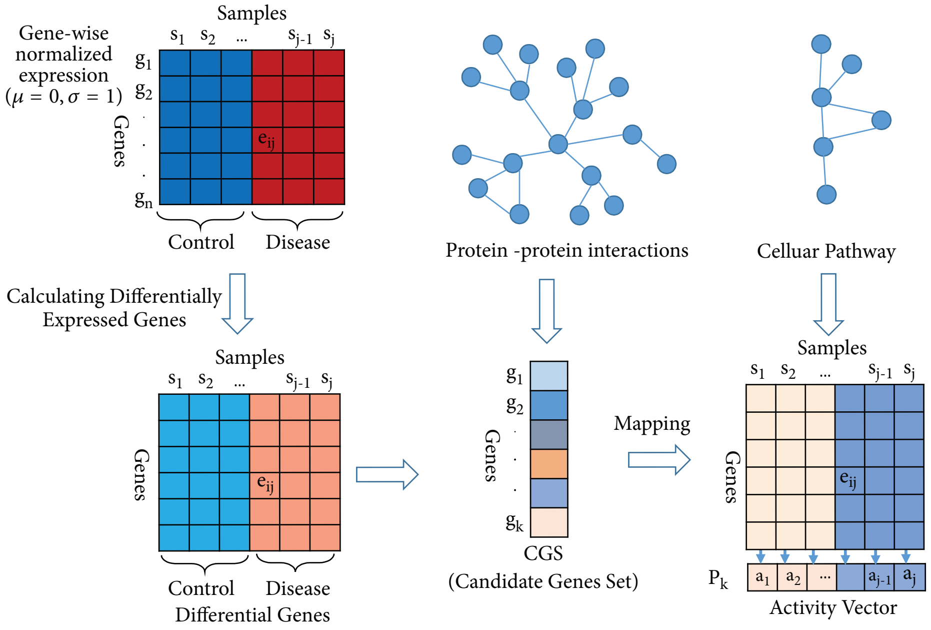
Gene-wise normalized expression ($\mu = 0, \sigma = 1$)
Samples
s_1 s_2 ... s_{j-1} s_j
g_1
g_2
.
.
.
g_n
Genes
e_{ij}
Control
Disease
Protein -protein interactions
Celluar Pathway
Calculating Differentially Expressed Genes
Samples
s_1 s_2 ... s_{j-1} s_j
Genes
e_{ij}
Control
Disease
Differential Genes
g_1
g_2
.
.
.
g_k
Genes
CGS
(Candidate Genes Set)
Mapping
Samples
s_1 s_2 ... s_{j-1} s_j
Genes
e_{ij}
P_k
a_1 a_2 ... a_{j-1} a_j
Activity Vector

Figure 2: Project the candidate gene to the pathway.

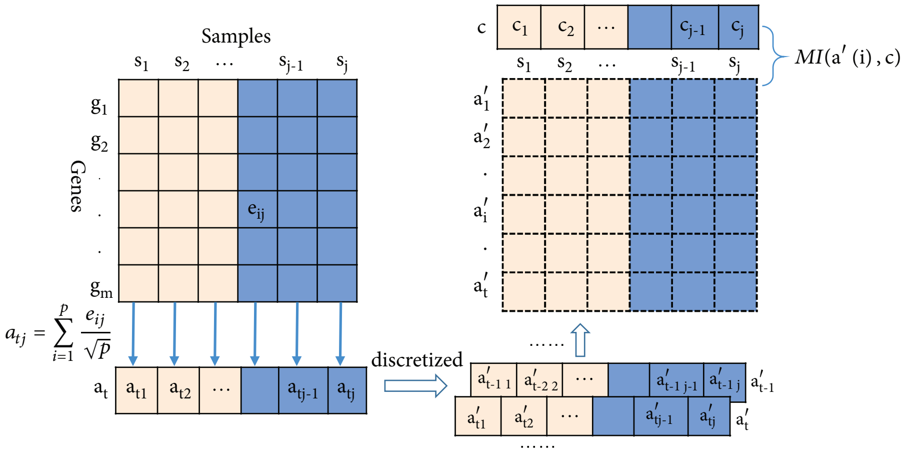

FIGURE 3: Calculation of mutual information.

(2) We compute differential expression genes in samples and filter through PPI network to get the candidate gene set (CGS).

(3) The biological signaling pathways are obtained from database MSigDB c2 (canonical pathway) which includes 1329 sets of biological metabolism and signaling pathways. The candidate genes are projected into the pathway, calculating the active score of each pathway.

(4) In the GEO data, we divide the samples into disease and control parts and normalize gene expression profiles of the samples and then calculate to get the differentially expressed genes. We map differential gene to the PPI network to obtain the candidate gene set (CGS). The activity scores are calculated according to the expression value of candidate gene set (CGS) on each sample. Gene expression value of gene in biological pathway for every sample constitutes an expression value matrix; each column represents a sample and each row represents a gene. A biological pathway includes p genes whose expression values can form a matrix of p rows. Each biological path corresponds to an activity vector, and the dimension of the vector is the number of samples; that is, in each sample j, an activity score can be calculated. The calculation formula is as follows:

$$a_{tj} = \sum_{i=1}^{p} \frac{e_{ij}}{\sqrt{p}} \tag{1}$$

a_{tj} represents the activity score of the t biological pathway in the j sample, and e_{ij} represents the gene expression value of the i gene in the t biological pathway of the sample j, and p is the number of genes in the biological pathway. After that, we get an activity vector of biological pathways; vector $[a_{t1}, a_{t2}, a_{t3}, \ldots a_{tm}]$ represents the activity score of the t biological pathway in m samples.

Next, a phenotypic vector is constructed based on the phenotypic labels of the samples, and then the mutual information between the activity vectors and the phenotypic vectors of the samples is computed by combining the activity vectors of each biological pathway.

2.2. Mutual Information With Phenotype. Mutual information (MI) is a commonly used method of information measurement in information theory. In [8], the rationale behind using MI to classify cancer patients is explained, and the processing is shown in Figure 3. By calculating the mutual information between the activity vectors and the phenotypic vectors of biological pathways, the correlations between the two vectors are measured, that is, the influence of a biological pathway on the phenotype of the disease.

Constructing a phenotype vector based on the phenotype of the sample, $[c1, c2, c3, \ldots cm]$, the phenotypic vector is a zero-one vector, and if the sample is tumor, the corresponding value is one, otherwise zero.

Using $a(i)$ to indicate the activity score of the i biological pathway on each sample, $a(i){=}[a_{i1}, a_{i2}, a_{i3}, \ldots a_{im}]$. c is used to represent the phenotypic vectors of m samples, $c = [c_1, c_2, c_3, \ldots c_m]$. So the correlation between biological pathways and disease phenotype $S(i)$ can be represented by mutual information $MI(a'(i), c)$ between $a'(i)$ vector and c vector. a' is a discretized form of a. The formula is as follows:

$$\begin{aligned} M(i) &= MI\left(a'(i), c\right) \\ &= \sum_{x \in a'} \sum_{y \in c} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \end{aligned} \tag{2}$$

The activity score a is discretized into $\lfloor \log 2(\mathit{sum\ of\ smaples}) + 1 \rfloor$ equally spaced bins to obtain a', respectively, $p(x, y)$ is the joint probability density function of a' and c, and $p(x)$ and $p(y)$ are the marginal probability density function of a' and c.

3. Constructing Pathway Network Related To Disease

The interaction of the pathway can be represented as a network which was constructed as follows.

Let $G(V, E)$ comprise a set V of pathways and a set E denote the weighted pathway-pathway interaction network with $E \in V * V$. Here, we use similarity to define weight of the network. $V = \{p_1, p_2 \ldots p_n\}$, $n = |V|$ is the number of pathways. Matrix A represents the weighted $n \times n$ adjacency matrix of G, where w_{ij} denotes the weight of the edge connecting pathway p_i to p_j and w_{ij} is calculated by the similarity of p_i and p_j. Supposing two pathways $p_1$and p_2, p_1, p_2 is a set of genes. $p_1 = \{g_{11}, g_{12} \ldots g_{1n}\}$, and $p_2 = \{g_{21}, g_{22} \ldots g_{2m}\}$.The similarity of p_1 and p_2 can be calculated as weight of network. The similarity between the two pathways is the ratio of the sum of the similarities of all similar genes to the sum of the two pathway's elements. The formula is shown in

$$sim_{pi,pj} = \frac{\sum_{g_{ix\in pi}, g_{jy\in pj}} sim\left(g_{ix}, g_{jy}\right)}{M + N} \tag{3}$$

$M = |p_i|$ is the number of genes on p_i, and $N = |p_j|$ is the number of genes on p_j.

By this formula, the similarity values of each pair of candidate pathways can be calculated, and the pathway network can be constructed to identify the critical pathway.

We use GOSemSim package of R language [9] to calculate the semantic similarity ($sim(g_{ix}, g_{jy})$) between genes, including molecular function similarity (MF), biological processes similarity (BP), and Cellular Component (CC) similarity. Since a pathway contains multiple genes, we calculate the similarity of genes on the pathway as the similarity between the pathways to obtain the pathway similarity matrix $sim_{pi,pj}$.

4. Identifying Significant Pathways of Pathway Network

Based on the biological pathway information and the similarity information of genes in the pathway, we constructed the biological pathway interaction network. In the network, each node represents a biological pathway, and if the two pathways contain differentially expressed genes, then they are connected. And the similarity of the differential genes on the two pathways is used as the similarity of the two pathways.

After constructing the biological pathway network, we hope to find biological pathways related to cancer phenotype. We use the random walk algorithm combined with phenotypic information to find the key nodes in the pathway network. The GeneRank method is a sorting algorithm proposed by Morrison et al. [10]. In our method, the initial node order is determined by the value of mutual information MI, and the transfer matrix is obtained by the pathway similarity matrix ($sim_{pi,pj}$). According to the random walk GeneRank algorithm, starting from the initial node, the final stable pathway node sequence is obtained by iterative calculation of the transfer matrix. This approach considers both the pathway and phenotype information and the semantic similarity between pathways, thus avoiding the node as isolated individuals and ignoring the important nodes that are highly correlated with other nodes. Therefore, this method can find potential pathway nodes.

We let set $P = \{p_{1,}p_{2,}, \ldots, p_N\}$ represent n nodes in the pathway network. According to whether there is a similarity relationship between the two pathway nodes, the adjacency matrix $w(i, j)$ can be obtained by a similar matrix $sim_{pi,pj}$ according to a threshold θ, as shown in

$$w(i, j) = \begin{cases} 1 & \text{if } \mathrm{sim}\left(\mathrm{pi}, \mathrm{pj}\right) \geq \theta \\ 0 & others \end{cases} \tag{4}$$

Sort the pathway set P according to the mutual information (MI) of pathway and phenotype. The initial row rank of P is obtained, which is denoted as x, $ex = (ex_{1,}ex_{2,}, \ldots, ex_N)$.

According to the definition, W is a symmetric matrix, $w_{i,j} = W(i, j) = W(j, i) = w_{j,i}$. According to graph theory, the degree of the i node is equal to the sum of the elements of row i of matrix w and is expressed in $\deg_i$, as shown in

$$deg_i = \sum_{j=1}^{n} w_{i,j} = \sum_{j=1}^{n} w_{j,i} \tag{5}$$

d is the damping coefficient, and the closer the value of d is to 0, the greater the impact of the mutual information is on the node sorting; on the contrary, if d is closer to 1, node sorting is more affected by the similarity. In our algorithm, we initialized d to 0.85.

The algorithm steps are as follows:

Input: initialization vector $ex = (ex_{1,}ex_{2,}, \ldots, ex_N)$; adjacency matrix w; parameter d.

ε is the error value; max is the maximum iterations.

Output: r;

(1) Data Preprocessing:

$$\begin{aligned} r^{[0]} &= \frac{ex}{\|ex\|_1}; \\ r^{[n]} &= \left(\mathrm{r}_1^{[n]}, \mathrm{r}_2^{[n]}, \ldots, \mathrm{r}_N^{[n]}\right); \quad n = 0, 1, 2, \ldots. \end{aligned} \tag{6}$$

(2) Iteration:

$$\begin{aligned} \mathrm{r}_j^{[n]} &= (1 - d)\, ex_j + d \sum_{i=1}^{N} \frac{w_{ij} r_i^{[n-1]}}{deg_i}, \quad 1 \leq j \leq N; \\ res^{[n]} &= \left\| r^{[n]} - r^{[n-1]} \right\|_1; \end{aligned} \tag{7}$$

(3) Stop Condition

If $res^{[n]} \leq \varepsilon\ or\ n \geq max\ , Stop$;

Else goto (2);

Return r=$r^{[n]}$;

5. Experiments and Results

5.1. Data. The Breast Cancer Dataset was downloaded from GEO (Gene Expression Omnibus) website (https://www.ncbi.nlm.nih.gov/geo/), including GSE33447 [11], GSE9309, GSE15852[12], GSE5364 [13]. and GSE20437 [14]. The dataset consists of 484 samples obtained from comparing 387 breast cancer samples with 97 normal samples, as shown in Table 1.

The gene expression profiles of 369 cases of breast cancer and 73 cases of normal breast tissues were obtained, and the differentially expressed genes were analyzed, PPI network was obtained from the Human Protein Reference

Table 1: Breast Cancer Data Set.

DataSet	Normal	Tumor
GSE 9309	9	132
GSE 15852	43	43
GSE 5364	13	186
GSE 33447	8	8
GSE 20437	24	18

Table 2: Top 15 pathways identified with our method.

Rank	Pathway Name	gene	p-value
1	KEGG_FATTY_ACID_METABOLISM	42	5.76e-12
2	KEGG_STARCH_AND_SUCROSE_METABOLISM	52	0.0047780
3	KEGG_SPLICEOSOME	128	3.63e-12
4	KEGG_PPAR_SIGNALING_PATHWAY	69	8.64e-21
5	KEGG_P53_SIGNALING_PATHWAY	69	0.0005045
6	KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	68	2.23e-09
7	PID_SHP2_PATHWAY	58	2.40e-05
8	PID_BARD1_PATHWAY	29	0.0010373
9	REACTOME_MRNA_3_END_PROCESSING	36	0.0137007
10	REACTOME_METABOLISM_OF_NON_CODING_RNA	49	0.0008123
11	REACTOME_CELL_CYCLE	421	1.02e-08
12	REACTOME_SIGNALING_BY_BMP	23	0.0247461
13	REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	54	0.0426007
14	REACTOME_CELL_CYCLE_MITOTIC	325	9.35e-07
15	REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	23	0.0127461

Database (http://www.hprd.org/) [15]. The pathways were downloaded from the Molecular Signatures Database website (http://software.broadinstitute.org/gsea/msigdb) [16]. The database mainly collects gene set which was annotated with certain biological functions. We chose C2 gene sets (canonical pathways, 1392 gene sets); the set of genes is derived from several major biological pathway databases, including BioCarta [17], KEGG [18], and Reactome [19] databases.

To identify the significance of the given pathway, first, we computed differential expression genes of the sample datasets (GSE9309, GSE15852) and then dealt with the PPI data. PPI network was mapped into the differential expression gene; we obtained candidate gene set (CGS). Secondly, the candidate genes are mapped into the pathways, calculating the active score of each pathway, and then the mutual information is computed by combining the activity vectors of each biological pathway. Finally, we calculate the similarity of the pathway to obtain pathway network, and the random walk algorithm combined with phenotypic information was used to find significance node of the pathway network.

To provide a more comprehensive understanding of the proposed method, we discuss the method from the following aspects separately.

5.2. The Results of Pathway Recognition. According to the above description, the rank of each pathway is the degree of relevance between the given pathway and the corresponding disease. The rank is calculated by algorithm (see in Section 4). In this study, the significance of the pathway was tested by hypergeometric distribution of each pathway with annotated differential genes using formula (8). The top 15 pathways with p-value are shown in Table 2; we selected part of the interaction pathways, as shown in Figure 4.

$$p - value = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{U-M}{N-i}}{\binom{U}{N}} \tag{8}$$

U—the number of genes in the human genome;

N—the number of differential genes;

M—the number of genes in the pathway;

x—the number of differentially expressed genes in this pathway.

The top 1 pathway is KEGG_FATTY_ACID_METABOLISM; the pathway is supported by [20–22], and the KEGG_SPLICEOSOME [23–25], REACTOME_ METABOLISM_OF_NON_CODING_RNA [26, 27], and REACTOME_CELL_CYCLE [28] are all very important metabolic pathways, and we focus on the following important pathways.

One significant pathway identified by our method was P53_SIGNALING_PATH-WAY [29–31]. Currently, breast cancer is the most prevalent cancer diagnosed in women, with an estimated 1.8 million cases reported worldwide in 2013 [32, 33]. Radiation is commonly adopted as an adjuvant therapy for the management of breast cancer [34]. However,

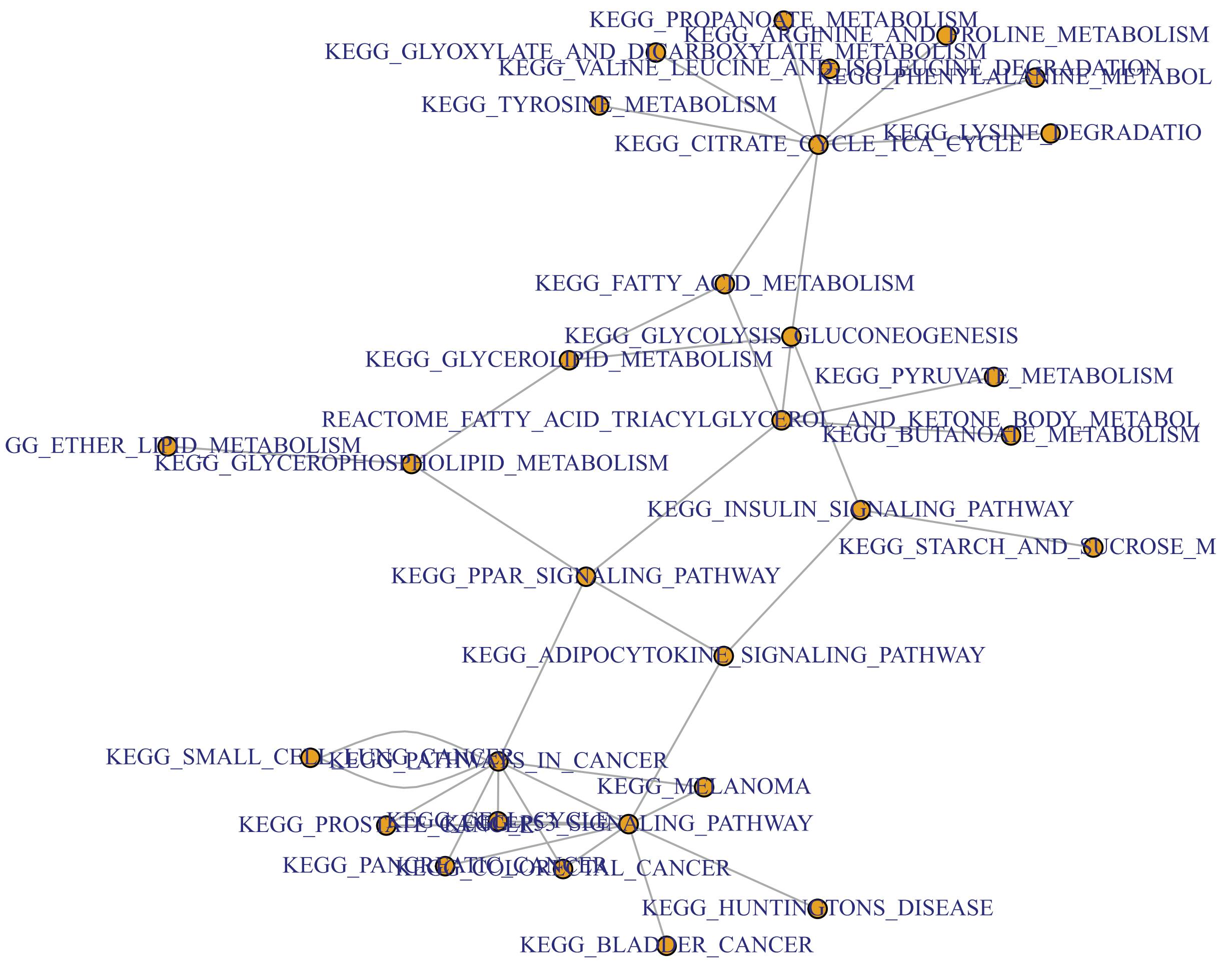
KEGG_PROPANOATE_METABOLISM
KEGG_ARGININE_AND_PROLINE_METABOLISM
KEGG_GLYOXYLATE_AND_DICARBOXYLATE_METABOLISM
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION
KEGG_PHENYLALANINE_METABOL
KEGG_TYROSINE_METABOLISM
KEGG_LYSINE_DEGRADATIO
KEGG_CITRATE_CYCLE_TCA_CYCLE
KEGG_FATTY_ACID_METABOLISM
KEGG_GLYCOLYSIS_GLUCONEOGENESIS
KEGG_GLYCEROLIPID_METABOLISM
KEGG_PYRUVATE_METABOLISM
REACTOME_FATTY_ACID_TRIACYLGLYCEROL_AND_KETONE_BODY_METABOL
KEGG_BUTANOATE_METABOLISM
GG_ETHER_LIPID_METABOLISM
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM
KEGG_INSULIN_SIGNALING_PATHWAY
KEGG_STARCH_AND_SUCROSE_M
KEGG_PPAR_SIGNALING_PATHWAY
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY
KEGG_SMALL_CELL_LUNG_CANCER
KEGG_PATHWAYS_IN_CANCER
KEGG_MELANOMA
KEGG_PROSTATE_CANCER
KEGG_CELL_CYCLE
KEGG_P53_SIGNALING_PATHWAY
KEGG_PANCREATIC_CANCER
KEGG_COLORECTAL_CANCER
KEGG_HUNTINGTONS_DISEASE
KEGG_BLADDER_CANCER

Figure 4: Interaction between disease pathways.

there is growing evidence that autophagy is induced by ionizing radiation, and this induction plays a crucial role in radiosensitivity [35, 36]. Furthermore, the regulatory effect of autophagy in radiation-induced cell death remains controversial, and the underlying molecular mechanisms remain to be fully characterized. The research of [37] shows that the p53/DRAM signaling pathway appears to contribute to radiation-induced autophagic cell death in MCF-7 breast cancer cells.

Another significant pathway was KEGG_PPAR_SIGNALING_PATHWAY. In [38], the authors analyzed six pathological complete response (pCR) patients and 25 patients with non-pCR; 300 probes (231 genes) were identified as differentially expressed between pCR and residual disease by the SAM program when the fold change was >2. The gene functional enrichment analysis revealed 15 prominent gene categories that were different between pCR and non-pCR patients, most notably the genes involved in the peroxisome proliferator-activated receptor (PPAR), DNA repair, and ER signal pathways and in the immune-related gene cluster; they believe that the PPAR pathway may be an important predictor of genes that are involved in the chemotherapy response.

The other pathway was PID_SHP2_PATHWAY; in [39], the researchers show a fundamental role for Src-homology 2 domain-containing phosphatase 2 (SHP2) in these processes in human epidermal growth factor receptor 2- (HER2-) positive and triple-negative breast cancers. Knockdown of SHP2 eradicated breast tumor-initiating cells in xenograft models, and SHP2 depletion also prevented invasion in three-dimensional cultures and in a transductal invasion assay in vivo. Notably, SHP2 knockdown in established breast tumors blocked their growth and reduced metastasis. Mechanistically, SHP2 activated stemness-associated transcription factors, including v-myc myelocytomatosis viral oncogene homolog (c-Myc) and zinc finger E-box binding homeobox 1 (ZEB1), which resulted in the repression of let-7 microRNA and the expression of a set of “SHP2 signature” genes. They found these genes to be simultaneously activated in a large subset of human primary breast tumors that are associated with invasive behavior and poor prognosis. These results provide new insights into the signaling cascades influencing tumor-initiating cells as well as a rationale for targeting SHP2 in breast cancer.

Moreover, we obtained top 15 pathways; their p-values are all less than 0.05. In order to test the effectiveness of this

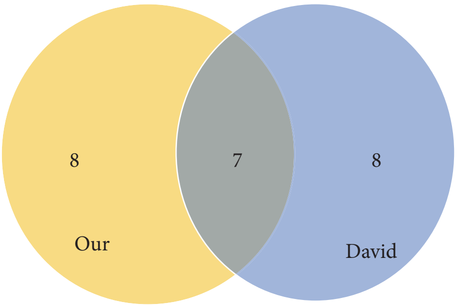

Figure 5: DAVID method and our method to identify the risk pathway.

Table 3: Common pathways with our method and David.

Pathway Name	p-value
KEGG_FATTY_ACID_METABOLISM	5.76e-12
KEGG_SPLICEOSOME	3.63e-12
KEGG_PPAR_SIGNALING_PATHWAY	8.64e-21
KEGG_P53_SIGNALING_PATHWAY	0.0005045
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	2.23e-09
REACTOME_CELL_CYCLE	1.02e-08
REACTOME_CELL_CYCLE_MITOTIC	9.35e-07

method, we compare with the most used DAVID software. The identified top 15 pathways are compared with DAVID method and our method to identify the risk pathway of breast cancer as shown in Figure 5; the common pathways are shown in Table 3.

DAVID is a widely used and approved method. According to the pathway set we have identified (in Tables 2 and 5). Pathways identified by our method and DAVID method have obvious intersection (in Figure 5); the p-value of the pathway is also significant. In addition, compared with DAVID, we also found some pathways which DAVID did not find. In Table 2, PID_BARD1_PATHWAY [40–45], PID_SHP2_PATHWAY [39], REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM [46], and REACTOME_METABOLISM_OF_NON_CODING_RNA [46, 47] are the pathways that our method identified but DAVID did not. The relationship between them and breast cancer has been evaluated in detail and is affirmative in the corresponding literature.

So, according to the above analysis, it can be concluded that our proposed method is effective in identifying the important pathways of the complex diseases.

5.3. Test on Dataset

5.3.1. Results and Analysis. To estimate the classification performance, firstly we prepared our dataset (GSE9309, GSE15852) and took 80 genes in the selected pathways as features, and SVM [48] is employed to classify the selected samples. Next, a 10-fold cross-validation was used to train and test SVM. The above experiment was repeated 100 times; the average value of the 100-time calculation is taken as the final result. In order to evaluate our method, we compared the classification results of cancer and normal samples with the commonly used methods based on differential expression. We choose the T-test method, and the T-test can be used to test whether the means of two independent normal distribution samples are equal. For gene expression data, we can test whether there is a significant difference in the expression of a gene between different phenotypes, that is, to identify differentially expressed genes through tests. The results obtained by the T-test method are compared with our method according to the genes contained in the biological pathway, and the same number of genes is sorted according to the order. The comparison results show that the method is superior to the T-test method, and the experimental results show that our proposed method is more effective in distinguishing cancer from normal samples, and the results are shown in Figure 6.

We applied the feature gene set to the independent gene expression datasets (GSE5364, GSE33447, and GSE20437). This set of data is not related to the data previously used, and they are independent of each other. In order to evaluate our proposed method objectively, we used the same number of biological pathways or gene markers to compare the results. The biological pathway markers obtained by our method have a good discrimination in the dataset, higher than 0.6, close to 0.7, and the results are shown in Figure 7. However, the AUC of the currently reported prognostic models on independent datasets is very difficult to reach 0.7 [49], which shows the superiority of our method.

5.3.2. Test on Other Datasets. Secondly, in order to test the robustness of our method, we apply the method in this article to four gastric cancer gene expression datasets. The dataset was also downloaded from GEO, which includes GSE63089 [50], GSE56807 [51], GSE33335 [52], and GSE19826 [53]. PPI data was obtained from STRING (https://stringdb-static.org/); the sources of other data are the same. The dataset consists of 177 samples obtained from comparing 87 tumor samples with 90 normal samples, which is shown in Table 4; the top 15 pathways with p-value are shown in Table 5.

We used our method to test on the gastric cancer dataset, and we selected the same number of biological pathways and gene markers. Compared with the T-test, our method has also achieved a higher or comparable value. We used GSE63089 and GSE56807 as the training set, with GSE33335 and GSE19826 as the test set. A 10-fold cross-validation was used to train and test SVM. The above experiment was repeated 50 times; the average value is taken as the final result. Our method also has good discrimination in this dataset, higher than 0.6 and better than T-test; the results on gastric cancer data are shown in Figure 8.

5.4. Conclusions. Complex diseases, especially cancer, are extremely harmful to human health. Therefore, the identification of cancer markers is the key of the study. Pathway analysis combined with multiple types of high-throughput data reflects the biological processes more clearly. Therefore, pathway-based complex disease analysis method has become

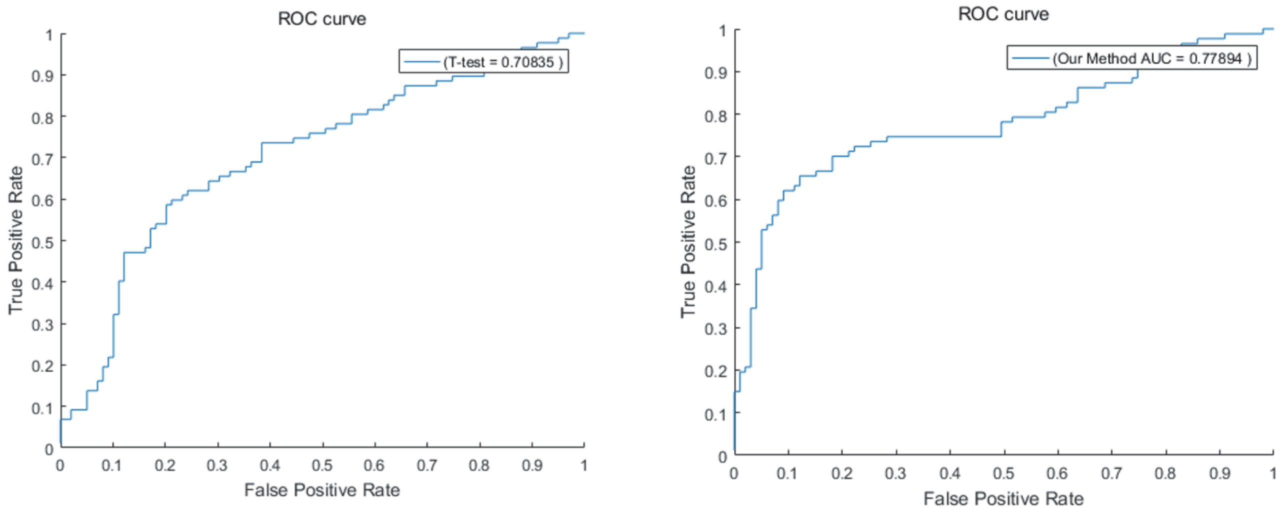

Figure 6: Test on dataset.

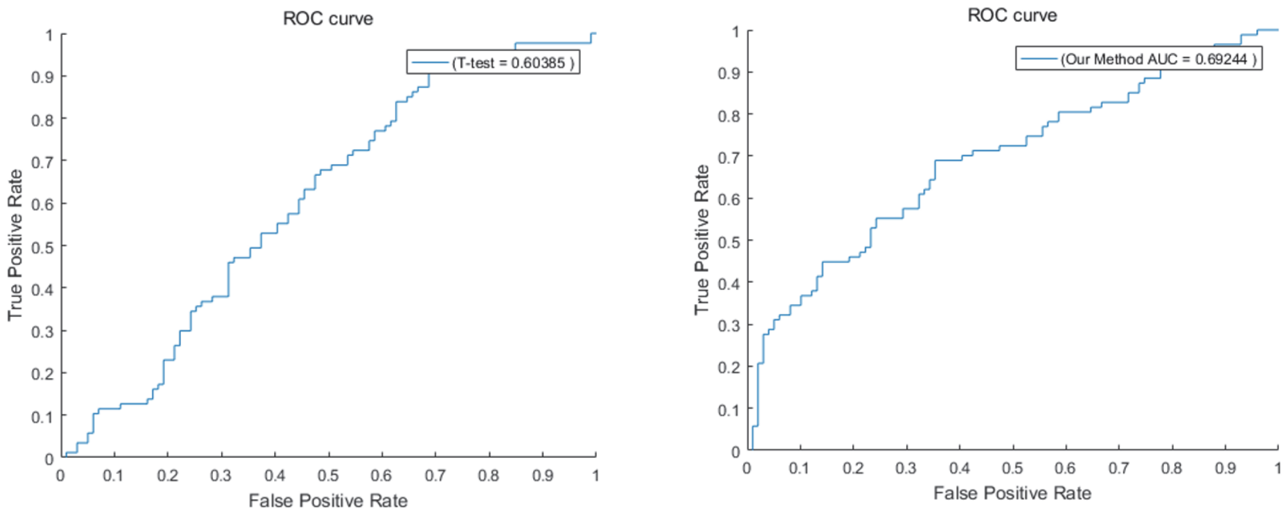

Figure 7: Test on independent datasets.

a hot research topic. Unlike previously available pathway analysis methods, we have considered not only the genes interaction of the pathway, but also the interaction between the pathways.

In this paper, we proposed a new approach to consider the correlation between biological pathways and establish a biological pathway interaction network. Then, the GeneRank algorithm based on random walk and the mutual information of phenotype were used to select cancer related biological pathways. Finally, we use the support vector machine and feature selection method to apply to cancer datasets. The results show that our method achieves better results than T-test method. In addition, the validation in the independent dataset and the functional analysis of the biological pathway indicate that the pathway we identified as a biological marker of disease is more accurate and reliable. We will employ more datasets to assess the validity of our approach in future research.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study is supported by the Natural Science Foundation of Hubei Province of China (Program no. 2015CFB524 and Program no. 2016CKB705), the Fundamental Research Funds for the Central Universities (Program no. 2015BQ023 and Program no. 2014QC008), and the Students Research Fund

Table 4: Gastric Cancer Data Set.

DataSet	Normal	Tumor
GSE 63089	45	45
GSE 56807	5	5
GSE 33335	25	25
GSE 19826	15	12

Table 5: Top 15 pathways identified with our method in Gastric Cancer Data Set.

Rank	Pathway Name	gene	p-value
1	KEGG_DNA_REPLICATION	36	1.50e-20
2	KEGG_PROTEASOME	48	9.07e-05
3	KEGG_ARGININE_AND_PROLINE_METABOLISM	54	0.000489
4	KEGG_GLYCEROLIPID_METABOLISM	49	5.12e-05
5	KEGG_PURINE_METABOLISM	159	1.68e-12
6	KEGG_PATHWAYS_IN_CANCER	328	2.52e-09
7	KEGG_SPLICEOSOME	128	0.0003348
8	KEGG_NUCLEOTIDE_EXCISION_REPAIR	44	0.0001406
9	KEGG_MELANOGENESIS	102	0.0481885
10	KEGG_RNA_DEGRADATION	59	0.0003626
11	KEGG_MAPK_SIGNALING_PATHWAY	267	0.0014066
12	KEGG_GLYCEROLIPID_METABOLISM	49	5.119e-05
13	KEGG_DRUG_METABOLISM_CYTOCHROME_P450	72	0.0115133
14	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	267	0.0183417
15	KEGG_PROSTATE_CANCER	89	0.0274642

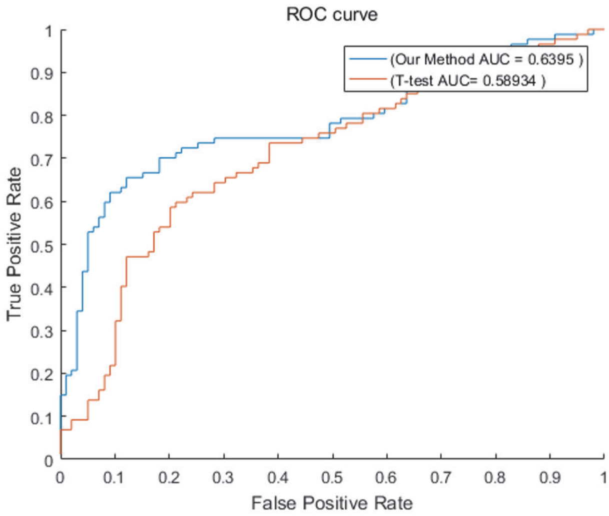

Figure 8: Test on independent gastric cancer dataset.

(SRF) of Huazhong Agricultural University (Program no. 2017400).

References

[1] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: A universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, 2017.

[2] U. D. Akavia, O. Litvin, J. Kim et al., "An integrated approach to uncover drivers of cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, 2010.

[3] Q. Zhang, J. Li, H. Xue, L. Kong, and Y. Wang, "Network-based methods for identifying critical pathways of complex diseases: A survey," *Molecular BioSystems*, vol. 12, no. 4, pp. 1082–1089, 2016.

[4] C. S. Greene and B. F. Voight, "Pathway and network-based strategies to translate genetic discoveries into effective therapies," *Human Molecular Genetics*, vol. 25, no. 2, pp. R94–R98, 2016.

[5] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[6] A. L. Tarca, S. Draghici, P. Khatri et al., "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[7] https://david.ncifcrf.gov/.

[8] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, no. 1, article 140, 2007.

[9] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: An R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, Article ID btq064, pp. 976–978, 2010.

[10] J. L. Morrison and etal., "GeneRank: Using search engine technology for the analysis of microarray experiments," *Bmc Bioinformatics*, vol. 6, pp. 1–14, 2005.

[11] Z.-Q. Lian, Q. Wang, W.-P. Li, A.-Q. Zhang, and L. Wu, "Screening of significantly hypermethylated genes in breast cancer using microarray-based methylated-CpG island recovery assay and identification of their expression levels," *International Journal of Oncology*, vol. 41, no. 2, pp. 629–638, 2012.

[12] I. B. Pau Ni, "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context," *Pathology Research & Practice*, vol. 206, pp. 223–228, 2010.

[13] K. Yu, K. Ganesan, L. K. Tan et al., "A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000129, 2008.

[14] K. Graham, A. De Las Morenas, A. Tripathi et al., "Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile," *British Journal of Cancer*, vol. 102, no. 8, pp. 1284–1293, 2010.

[15] B. Liu and B. Hu, "HPRD: a high performance RDF database," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 25, no. 2, pp. 123–133, 2010.

[16] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739-1740, 2011.

[17] http://www.biocarta.com/.

[18] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, vol. 38, no. 1, pp. D355–D360, 2009.

[19] E. Schmidt, E. Birney, D. Croft et al., "Reactome – A Knowledgebase of Biological Pathways," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, vol. 4277 of *Lecture Notes in Computer Science*, pp. 710–719, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[20] A. Huqi, "Cancer and inhibition of fatty acid oxidation," *Heart and Metabolism*, no. 51, pp. 27–30, 2011.

[21] W. Zhou, Y. Tu, P. J. Simpson, and F. P. Kuhajda, "Malonyl-CoA decarboxylase inhibition is selectively cytotoxic to human breast cancer cells," *Oncogene*, vol. 28, no. 33, pp. 2979–2987, 2009.

[22] J. N. Thupari, M. L. Pinn, and F. P. Kuhajda, "Fatty acid synthase inhibition in human breast cancer cells leads to malonyl-CoA-induced inhibition of fatty acid oxidation and cytotoxicity," *Biochemical and Biophysical Research Communications*, vol. 285, no. 2, pp. 217–223, 2001.

[23] M. C. Wahl, C. L. Will, and R. Lührmann, "The spliceosome: design principles of a dynamic RNP machine," *Cell*, vol. 136, no. 4, pp. 701–718, 2009.

[24] H. Le Hir, E. Izaurralde, L. E. Maquat, and M. J. Moore, "The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions," *EMBO Journal*, vol. 19, no. 24, pp. 6860–6869, 2000.

[25] J. P. Staley and C. Guthrie, "Mechanical devices of the spliceosome: Motors, clocks, springs, and things," *Cell*, vol. 92, no. 3, pp. 315–326, 1998.

[26] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[27] S. Sarkar, H. Dubaybo, S. Ali, P. Goncalves, and S. L. Kollepara, "Long non-coding RNAs (lncRNAs) and viral infections," *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, vol. 3, pp. 465–477, 2013.

[28] T. Mazza and M. Cavaliere, "Cell Cycle and Tumor Growth in Membrane Systems with Peripheral Proteins," *Electronic Notes in Theoretical Computer Science*, vol. 227, no. C, pp. 127–141, 2009.

[29] Y. N. Lee, Y. G. Park, Y. H. Choi, and Y. S. Cho, "CRE-transcription factor decoy oligonucleotide inhibition of MCF-7 breast cancer cells: Cross-talk with p53 signaling pathway," *Biochemistry*, vol. 39, no. 16, pp. 4863–4868, 2000.

[30] O. E. C. Mg, "Differential regulation of signal transduction pathways in wild type and mutated p53 breast cancer epithelial cells by copper and zinc," *Archives of Biochemistry & Biophysics*, vol. 423, no. 2, pp. 351–361, 2004.

[31] J. H. Seo, "PEGylated conjugated linoleic acid stimulation of apoptosis via a p53-mediated signaling pathway in MCF-7 breast cancer cells," *European Journal of Pharmaceutics & Biopharmaceutics Official Journal of Arbeitsgemeinschaft Für Pharmazeutische Verfahrenstechnik E V*, vol. 70, no. 621, 2008.

[32] K. N. Dalby, I. Tekedereli, G. Lopez-Berestein, and B. Ozpolat, "Targeting the prodeath and prosurvival functions of autophagy as novel therapeutic strategies in cancer," *Autophagy*, vol. 6, no. 3, pp. 322–329, 2010.

[33] C. Fitzmaurice, D. Dicker, A. Pain et al., "The global burden of cancer 2013," *JAMA Oncology*, vol. 1, pp. 505–527, 2015.

[34] P. G. Tsoutsou, M. I. Koukourakis, D. Azria, and Y. Belkacémi, "Optimal timing for adjuvant radiation therapy in breast cancer. A comprehensive review and perspectives," *Critical Review in Oncology/Hematology*, vol. 71, no. 2, pp. 102–116, 2009.

[35] Y. Yang, "Autophagy and its function in radiosensitivity," *Tumour Biology the Journal of the International Society for Oncodevelopmental Biology & Medicine*, vol. 36, pp. 1–9, 2015.

[36] D. A. Gewirtz, "The four faces of autophagy: implications for cancer therapy," *Cancer Research*, vol. 74, no. 3, pp. 647–651, 2014.

[37] L. Cui, Z. Song, B. Liang, L. Jia, S. Ma, and X. Liu, "Radiation induces autophagic cell death via the p53/DRAM signaling pathway in breast cancer cells," *Oncology Reports*, vol. 35, no. 6, pp. 3639–3647, 2016.

[38] Y. Z. Chen, "PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy," *Cancer Chemotherapy & Pharmacology*, vol. 70, 2012.

[39] X. Wang, "Abstract 2150: The role of SHP2 in HER2+ breast cancer," *Cancer Research*, vol. 72, no. 8, pp. 2150-2150, 2014.

[40] M. Fabbro, K. Savage, K. Hobson et al., "BRCA1-BARD1 complexes are required for P53Ser-15 phosphorylation and a G1/S arrest following ionizing radiation-induced DNA damage," *The Journal of Biological Chemistry*, vol. 279, no. 30, pp. 31251–31258, 2004.

[41] G. Birrane, A. K. Varma, A. Soni, and J. A. A. Ladias, "Crystal structure of the BARD1 BRCT domains," *Biochemistry*, vol. 46, no. 26, pp. 7706–7712, 2007.

[42] R. Baer and T. Ludwig, "The BRCA1/BARD1 heterodimer, a tumor suppressor complex with ubiquitin E3 ligase activity," *Current Opinion in Genetics & Development*, vol. 12, no. 1, pp. 86–91, 2002.

[43] "Bard, BRCA1 associated RING domain 1".

[44] D. Fox III, I. Le Trong, P. Rajagopal, P. S. Brzovic, R. E. Stenkamp, and R. E. Klevit, "Crystal structure of the BARD1 ankyrin repeat domain and its functional consequences," *The Journal of Biological Chemistry*, vol. 283, no. 30, pp. 21179–21186, 2008.

[45] I. Irminger-Finger, W.-C. Leung, J. Li et al., "Identification of BARD1 as mediator between proapoptotic stress and p53-dependent apoptosis," *Molecular Cell*, vol. 8, no. 6, pp. 1255–1266, 2001.

[46] P. Boyle and P. Boffetta, "Alcohol consumption and breast cancer risk," *Breast Cancer Research*, vol. 11, no. S3, 2009.

[47] P. Paci, T. Colombo, and L. Farina, "Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer," *BMC Systems Biology*, vol. 8, article 83, 2014.

[48] C. Chang and C. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[49] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel, "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context," *BMC Bioinformatics*, vol. 11, article no. 277, 2010.

[50] X. T. Zhang, Z. H. Ni, Z. P. Duan et al., "Overexpression of E2F mRNAs associated with gastric cancer progression identified by the transcription factor and miRNA co-regulatory network analysis," *PLoS ONE*, vol. 10, no. 2, article e0116979, 2015.

[51] J. Wang, Z. Ni, Z. Duan, G. Wang, and F. Li, "Altered expression of hypoxia-inducible factor-1α (HIF-1α) and its regulatory genes in gastric cancer tissues," *PLoS ONE*, vol. 9, no. 6, Article ID e99835, 2014.

[52] L. Cheng, P. Wang, S. Yang et al., "Identification of genes with a correlation between copy number and expression in gastric cancer," *BMC Medical Genomics*, vol. 5, article 14, 2012.

[53] Q. Wang, Y. G. Wen, D. P. Li et al., "Upregulated INHBA expression is associated with poor survival in gastric cancer," *Medical Oncology*, vol. 29, no. 1, pp. 77–83, 2012.