Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 836929, 8 pages
http://dx.doi.org/10.1155/2015/836929

Research Article

Module Based Differential Coexpression Analysis Method for Type 2 Diabetes

Lin Yuan,[1] Chun-Hou Zheng,[2] Jun-Feng Xia,[3] and De-Shuang Huang[1]

[1]*School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2]*College of Electrical Engineering and Automation, Anhui University, Hefei 230601, China*
[3]*Institute of Health Sciences, Anhui University, Hefei 230601, China*

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 4 December 2014; Accepted 29 December 2014

Academic Editor: Fang-Xiang Wu

More and more studies have shown that many complex diseases are contributed jointly by alterations of numerous genes. Genes often coordinate together as a functional biological pathway or network and are highly correlated. Differential coexpression analysis, as a more comprehensive technique to the differential expression analysis, was raised to research gene regulatory networks and biological pathways of phenotypic changes through measuring gene correlation changes between disease and normal conditions. In this paper, we propose a gene differential coexpression analysis algorithm in the level of gene sets and apply the algorithm to a publicly available type 2 diabetes (T2D) expression dataset. Firstly, we calculate coexpression biweight midcorrelation coefficients between all gene pairs. Then, we select informative correlation pairs using the "differential coexpression threshold" strategy. Finally, we identify the differential coexpression gene modules using maximum clique concept and k-clique algorithm. We apply the proposed differential coexpression analysis method on simulated data and T2D data. Two differential coexpression gene modules about T2D were detected, which should be useful for exploring the biological function of the related genes.

1. Introduction

DNA microarray has been widely used as measurement tools in gene expression data analysis [1–4]. Gene expression profiling data from DNA microarray can detect the expression levels of thousands of genes simultaneously, providing an effective way for mining disease-related genes and revealing information of the regulatory networks and biological pathways of genes. Currently, the analysis of gene expression data can be divided into three levels: first, analysis of the expression level of individual genes, determining its function based on gene expression level changes under different experimental conditions: for example, the tumor type specific genes are identified according to the significance of difference in gene expression using the statistical hypothesis testing analysis method; second, study of gene interaction and coregulation through the combination of genes and grouping; and, third, an attempt to deduce the potential gene regulatory networks mechanism and explain the observed gene expression data.

Among the microarray data analysis methods, gene differential expression analysis is one of the most widely used types of analysis for disease research. Gene differential expression analysis method selects differentially expressed genes according to expression change value of a single gene. In fact, gene expression value change between normal samples and disease samples can be used to present the possibility of the relation between gene and disease. However, the traditional pathogenicity genes selection methods based on gene expression data treat each gene individually and interaction between them is not considered. Actually, genes and their protein products do not perform their functions in isolation [5, 6], but in cooperation. Functional changes such as alteration in tumor cell growth process, energy metabolism, and immune activity were accompanied with coexpression changes. Differentially expressed genes selection methods often focus only on the size of the single genes and

the relationship of individual genes and disease, ignoring a plurality of pathogenic genes of the complex disease as a gene module with disease related, as well as within the module gene [7].

Differential coexpression analysis, as a more comprehensive technique to the differential expression analysis, was raised to research gene regulatory networks and biological pathways of phenotypic changes through measure gene correlation changes between disease and normal conditions. Differential coexpression genes are defined as genes whose correlated expression pattern differs between classes [8]. The gene coexpression changes between different conditions indicate gene regulatory pathways and networks associated with disease. In gene differential coexpression analysis, a pair of gene expression datasets under disease and normal conditions is transformed to a pair of coexpression matrix in which links represent transcriptionally correlated gene pairs [5]. Until now, methods for differential coexpression analysis of gene expression data have been extensively researched, and multiple algorithms have been developed and tested [9–12]. In those gene differential coexpression analysis methods, the most common choice of similarity measurement is Pearson's correlation coefficients. However, Pearson's correlation is sensitive to outliers. So biweight midcorrelation (bicor) is considered to be a good alternative to Pearson's correlation since it is more robust to outliers [13].

In biomedical research, many complex diseases are contributed jointly by alterations of numerous genes; they often coordinate together as a functional biological pathway or network and are highly correlated. With recent interest of gene differential coexpression analysis in the gene network or module, gene module analysis has emerged as a novel holistic approach for microarray analysis. Somewhat large units, made up of genes, are more densely connected to each other than to the rest of the network, are often referred to as modules, and have been considered to be the essential structural units of real gene networks. There exists overlap among gene modules in large real networks.

Until now, there are many methods to find gene modules. For example, Butte and Kohane [14] proposed a systems-based approach called Entropy Minimization and Boolean Parsimony (EMBP) that identifies, directly from gene expression data, modules of genes that are jointly associated with disease. Kostka and Spang [15] used additive model to find differential coexpression gene modules. Prieto et al. [16] used altered expression based on improved additive model, optimal residual ratio, and minimum F-distribution to find differential coexpression gene modules. However, the microarray data contains a large number of genes; those methods need to search all gene expression data resulting in a large amount of computation; the process is very time-consuming even using optimized search algorithm.

The maximum clique analysis can avoid exhaustive search and quickly find maximum gene module with biological significance. The maximum clique problem (MCP) is a classical combinatorial optimization problem in graph theory. In 1957, Ross and Harary [17] first proposed the deterministic algorithm to solve the maximum clique problem. Since then some researchers had presented a variety of algorithms to solve this problem. The maximum clique problem is widely used in different areas, such as signal transmission, computer vision, and biological research. In this study, a gene coexpression network can be treated as a graph; gene is represented by vertex and coexpression relationship is represented by edge. We will use k-clique algorithm [18], which is an effective and deterministic method for uniquely identifying overlapping modules in large real networks. We first show some basic definitions. k-cliques, the central objects of k-clique algorithm investigation, are defined as complete (fully connected) subgraphs of k vertices. k-clique adjacency is as follows: two k-cliques are adjacent if they share some vertices. k-clique chain is as follows: a subgraph, which is the union of a sequence of adjacent k-cliques. We use k-clique algorithm to find gene cliques, and maximum clique concept is used to quickly find large gene modules which are made of k-clique chain. For the sake of convenience, we use the terms graph and community or network interchangeably, the former stressing the mathematical concept and the latter the application.

In this paper, we proposed a new approach for gene differential coexpression analysis in gene modules level based on combining biweight midcorrelation, differential coexpression threshold strategy, and maximum clique concept and k-clique analysis. Biweight midcorrelation measures the coexpression relationship between genes and the k-clique analysis with maximum clique concept quickly finds maximum disease-related module with biological significance. We use the approach to further investigate the gene module in order to gain insight into coexpression relationship between genes. The algorithm can find differential coexpression disease genes modules and global coexpression patterns are determined for type 2 diabetes expression dataset. As far as we know, no one has done this experiment.

The rest of the paper is organized as follows. Section 2 describes the methods proposed in this study. The biweight midcorrelation coefficients, "gene differential coexpression threshold" strategy, and threshold selection strategy are first presented, and the algorithm of k-clique is consequently given. Section 3 presents the experiment on simulated data and type 2 diabetes (T2D) in rats dataset. Section 4 concludes the paper and outlines directions of future work.

2. Methods

2.1. Biweight Midcorrelation for Differential Coexpression. Differential coexpression analysis usually requires the definition of "distance" or "similarity" between measured datasets, the most common choice being Pearson's correlation coefficients. However, Pearson's correlation coefficient is sensitive to outliers [13]. Biweight midcorrelation is considered to be a good alternative to Pearson's correlation since it is more robust to outliers. Example of a gene expression matrix is as follows:

$$\begin{bmatrix} & \text{Gene1} & \text{Gene2} & \cdots & \text{Gene}p \\ Z_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Z_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}. \quad (1)$$

For each sample Z_i, we measure expression levels for a set of genes, so X_{ij} is the measurement of the expression level of the jth gene for the ith sample, where $j = 1, \ldots, p$. The xth column vector of matrix represents gene expression profile of gene X. In order to define the biweight midcorrelation (bicor) [13] of two numeric vectors $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$, we first define u_i, v_i with $i = 1, \ldots, m$:

$$u_i = \frac{x_i - \operatorname{med}(x)}{9\operatorname{mad}(x)}, \qquad v_i = \frac{y_i - \operatorname{med}(y)}{9\operatorname{mad}(y)}, \tag{2}$$

where $\operatorname{med}(x)$ is the median of vector x, $\operatorname{mad}(x)$ is the median absolute deviation of vector x, $\operatorname{mad}(x)$ is the median of new numeric vector in which each number is absolute difference between original vector value and $\operatorname{med}(x)$; this leads us to the definition of $\operatorname{mad}(x)$ and weight w_i for x_i, which are

$$\operatorname{mad}(x) = \operatorname{med}\left(\left|x_i - \operatorname{med}(X)\right|\right), \qquad w_i^{(x)} = \left(1 - u_i^2\right)^2 I\left(1 - \left|u_i\right|\right), \tag{3}$$

where the indicator $I(1 - |u_i|)$ takes 1 if $1 - |u_i| > 0$ and 0 otherwise. Thus, the weight $w_i^{(x)}$ is close to 1 if x_i is close to $\operatorname{med}(x)$, approaches 0 when x_i differs from by nearly $9\operatorname{mad}(x)$, and is 0 if x_i differs from $\operatorname{med}(x)$ by more than $9\operatorname{mad}(x)$. An analogous weight $w_i^{(x)}$ can be defined for y_i. Given the weights, we can define biweight midcorrelation of x and y as

$$\operatorname{bicor}(x, y) = \sum_{i=1}^{m} \left(x_i - \operatorname{med}(x)\right) w_i^{(x)} \left(y_i - \operatorname{med}(y)\right) w_i^{(y)} \cdot \left(\sqrt{\sum_{j=1}^{m} \left[\left(x_j - \operatorname{med}(x)\right) w_j^{(x)}\right]^2} \cdot \sqrt{\sum_{k=1}^{m} \left[\left(y_k - \operatorname{med}(y)\right) w_k^{(y)}\right]^2} \right)^{-1}. \tag{4}$$

It should be noted that the equations of biweight midcorrelation do not involve an explicit identification of outliers, and all elements whose weight $w_i = 0$ can be considered outliers. The user can also set up the maximum allowed proportion of outliers using the argument "maxPOutliers"; the "maxPOutliers" is interpreted as the maximum proportion of low and high outliers separately. For the value of bicor from –1 to 1, –1 represents the maximum negative correlation and 1 represents the maximum positive correlation. Zero represents irrelevant correlation.

2.2. The "Differential Coexpression Threshold" Strategy. We used biweight midcorrelation to measure every pair of genes in the gene expression dataset and get a gene coexpression matrix. The gene coexpression matrix is a square and symmetric matrix P whose rows and columns correspond to the genes and whose element P_{ij} denotes the coexpression relationship between genes i and j. In this paper, we use A_{GN} which represents gene coexpression adjacency matrix in normal conditions and A_{GD} which represents gene coexpression adjacency matrix in disease condition. To find differential coexpression gene modules which are coexpressed in normal condition and not related to disease condition, we set two thresholds T_1 for adjacency matrix A_{GN} in normal condition and T_2 for adjacency matrix A_{GD} in disease condition. $A_{GN}(i, j)$ is set to 1 if value of $A_{GN}(i, j)$ is greater than or equal to T_1; otherwise, $A_{GN}(i, j)$ is set to 0 and $A_{GD}(i, j)$ is set to 1 if value of $A_{GD}(i, j)$ is less than or equal to T_2; otherwise, $A_{GD}(i, j)$ is set to 0. We integrated A_{GN} and A_{GD} into a matrix A_G after we had intersection of the corresponding elements of A_{GN} and A_{GD}. $A_G(i, j) = 1$ means coexpression value of gene i and gene j in A_{GN} is greater than or equal to T_1, and coexpression value of genes i and j in A_{GD} is less than or equal to T_2. $A_{GN}(i, j)$ also can be set to 1 if value of $A_{GN}(i, j)$ is less than or equal to T_1 and $A_{GD}(i, j)$ is set to 1 if value of $A_{GD}(i, j)$ is greater than or equal to T_2. The method is shown in (5).

With the above mentioned strategy, we also set $A_G(i, j) = 1$ if the absolute value of $A_{GN}(i, j)$ subtracting $A_{GD}(i, j)$ is greater than or equal to T_3 and the absolute value of $A_{GN}(i, j)$ is greater than or equal to the absolute value of $A_{GD}(i, j)$ simultaneously. This is a special type of coexpression change. In reality, coexpression reversal probably has biological significance. The coexpression reversal between normal condition and disease condition has advantage in disease. For example, the coexpression of *p53* and *Klf4* recently reported that the positive or negative correlation between these two genes determines the outcome of DNA damage, DNA repair, or apoptosis [19]. We believe that our attention to this special coexpression change will help to explore subtle mechanisms involved in genes transcriptional regulation. We excavated maximum cliques which have biological significance from A_G adjacency matrix to further investigate gene regulatory networks. Consider the following:

$$\begin{aligned}
&\text{if } A_{GN}(i, j) \geq T_1, \text{ then } A_{GN}(i, j) = 1,\\
&\text{else } A_{GN}(i, j) = 0;\\
&\text{if } A_{GD}(i, j) \leq T_2, \text{ then } A_{GD}(i, j) = 1,\\
&\text{else } A_{GD}(i, j) = 0;\\
&A_G(i, j) = A_{GN}(i, j) \cap A_{GD}(i, j)\\
&\text{if } A_{GN}(i, j) \leq T_1, \text{ then } A_{GN}(i, j) = 1,\\
&\text{else } A_{GN}(i, j) = 0;\\
&\text{if } A_{GD}(i, j) \geq T_2, \text{ then } A_{GD}(i, j) = 1,\\
&\text{else } A_{GD}(i, j) = 0;\\
&A_G(i, j) = A_{GN}(i, j) \cap A_{GD}(i, j).
\end{aligned} \tag{5}$$

2.3. The Threshold Selection Strategy. The two real value adjacency matrixes are transformed into a binary matrix which contains two elements 0 and 1 only. Choosing different thresholds will lead to different results; too large T_1 threshold or too small T_2 threshold will lead to small link number, low density clique, and lost biological significance cliques. On the other hand, too small T_1 or too large T_2 will lead to many overlapping cliques. They are not helpful for finding biological significance differential coexpression gene disease-related modules. In fact, how to choose a reasonable threshold in conversion process is a problem which needs to be further studied. Generally, the selection of the threshold can be based on the proportion of outliers in the figure or the density of graph. The outlier is the point which is not connected to any edges. The density is defined as the ratio of number of edges to the maximum possible number of edges in the graph. The density of clique is 1.

For gene expression data analysis, closely linked functional module is not the strict sense of maximum clique due to the lack of certain section. In this paper, we use density to measure approximation degree of functional module with gene differential coexpression clique, which may be having more biological significance.

2.4. The Maximum Clique Concept and k-Clique Algorithm. Graph theoretical concepts are useful for the description and analysis of interactions and relationships in biological systems. In gene coexpression graph, gene is represented by vertex and coexpression relationship by edge. $G = (V, E)$ is an arbitrary undirected and weighted graph unless otherwise specified in graph theoretical concepts. $V = \{1, 2, \ldots, n\}$ is the vertex set of G, and E is the edge set of G. For each vertex $i \in V$, a positive weight w_i is associated with i. $A_G = (a_{ij})_{n\times n}$ is the adjacency matrix of G, where $a_{ij} = 1$ if $(i, j) \in E$ is an edge of G, and $a_{ij} = 0$ if $(i, j) \notin E$. Genes and relationship between genes are represented by vertex and edge, respectively.

A graph $G = (V, E)$ is complete if all its vertices are pairwise adjacent; that is, for all $i, j \in V$, $(i, j) \in E$. A clique C is a subset of V such that $G(C)$ is complete. The maximum clique problem asks for a clique of maximum weight. An independent set (stable set and vertex packing) is a subset of V, whose elements are pairwise nonadjacent. The maximum independent set problem asks for an independent set of maximum cardinality. The size of a maximum independent set is the stability number of G (denoted by $\alpha(G)$). The maximum weight independent set problem asks for an independent set of maximum weight. A maximum clique means a clique which is a subset of the nodes in V in which every pair of nodes in the subset is joined by an edge and is not a proper subset of any other cliques [20].

In application, the identification of maximal cliques is often of limited interest since the requirement of complete connectivity is so restrictive. When dealing with imperfect systems or with experimental data, we may need to consider more general notions of cohesive subgroups. In this paper, we consider different notions of cohesive subgroups that include n-cliques, k-plexes, and λ-sets [18]. It is well known that the nodes of large real networks have a power law degree distribution [21]. Most real networks typically contain parts in which the nodes (units) are more highly connected to each other compared to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups, or modules [22–26], which have no widely accepted unique definition. The basic observation on which our modules definition relies is that a typical gene differential coexpression module consists of several complete (fully connected) subcliques that tend to share many of their nodes. To find meaningful communities, several basic requirements should be satisfied: it cannot be too restrictive, should be based on the density of links, is required to be local, should not yield any cut-node or cut link (whose removal would disjoin the community), and, of course, should allow overlaps. We employ the community definition specified above because none of the others in the literature satisfy all these requirements simultaneously [27–29].

k-clique algorithm for detecting gene differential coexpression modules in a network has been published in the paper [26]. k-clique algorithm is also named clique percolation method. The existing divisive and agglomerative methods recently used for large real networks have some disadvantages. Divisive methods cut the network into smaller and smaller pieces; each node is forced to remain in only one community and be separated from its other communities, most of which then necessarily fall apart and disappear [27, 30]. The agglomerative [31] method has the same problem. The k-clique algorithm has demonstrated the advantages over the divisive method and agglomerative method. In the algorithm, although the numerical determination of the full set of k-clique communities is a polynomial problem, the algorithm is exponential and significantly more efficient for the graphs corresponding to actual data. The k-clique algorithm first locates all cliques (maximal complete subgraphs) of the network and then identifies the communities by carrying out a standard component analysis of the clique-clique overlap matrix [28]. The k-clique algorithm uses the threshold probability $d(k)$ (critical point) of k-clique percolation to find all maximal complete subgraphs. The critical point is shown in (6), where N is the number of genes or vertex of graph:

$$d(k) = \frac{1}{[(k-1)N]^{1/(k-1)}}. \tag{6}$$

The k-clique algorithm gives two plausible choices to measure the size of the largest k-clique percolation cluster in (7) and (8). The most natural one, which we denote by N^*, is the number of vertices belonging to this cluster. ϕ is an order parameter associated with this choice as the relative size of that cluster:

$$\phi = \frac{N^*}{N}. \tag{7}$$

The other choice is the number L^* of k-cliques of the largest k-clique percolation cluster. The associated order parameter is again the relative size of this cluster:

$$\varphi = \frac{L^*}{L}, \tag{8}$$

where L denotes the total number of k-cliques in the graph. L can be estimated as

$$L \approx \binom{N}{k} d^{k(k-1)/2} \approx \frac{N^k}{k!} d^{k(k-1)/2}. \quad (9)$$

In this paper, we use the biweight midcorrelation for constructing binary networks. Two-condition coexpression adjacency networks can always be transformed into a binary one by ignoring any directionality in the links and keeping only those stronger than a threshold weight. Then, the concept of maximum clique and k-clique algorithm were used to find gene differential coexpression modules. We named the proposed method "BMKC" (biweight midcorrelation and k-clique algorithm) method. Changing the threshold is like changing the resolution with which the community structure is investigated: by increasing, the communities start to shrink and fall apart. A very similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated but, at the same time, also more cohesive. More details about k-clique algorithm can be found in [28, 32].

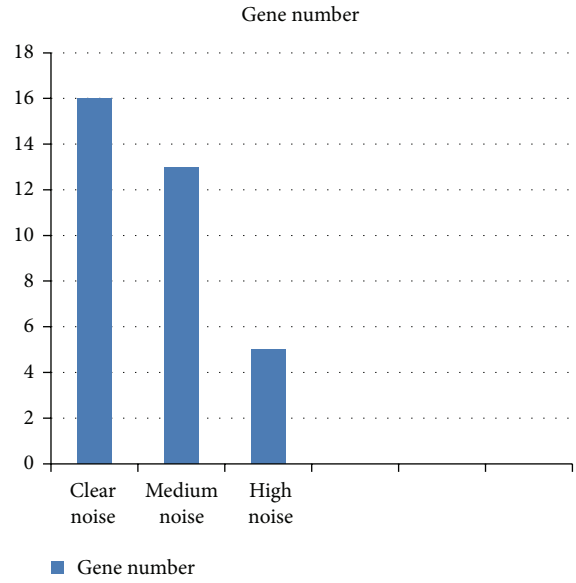

Figure 1: The column bar graph shows the effect of the noise parameter σ on the size of the gene group found by our algorithm.

3. Results

3.1. Experiment Result on Simulated Datasets. We first evaluate the algorithm in a supervised setting. We generate a control group of 30 samples and a disease group of another 30 samples, both consisting of 120 genes. For the control group, 20 coexpressed genes are sampled directly from the biweight midcorrelation. We focus on whether k-clique algorithm can find coexpression gene modules from the background of noise. We first draw a vector with 20 rows and a vector with 30 columns from a standard normal distribution. The actual expression levels are obtained by adding independent errors sampled from a normal distribution with mean zero and standard deviation (SD) σ. These 20 genes form the target pattern. We then hide them in 100 additional noise genes, which are sampled independent and identically distributed (i.i.d.) from a standard normal distribution. The disease group is simulated by 120 independent noise genes drawn from a standard normal only.

In the above setting, we use SD σ to tune the strength of the signal resulting from the 20 coexpressed genes. To observe its effect in detail, we use three different values: for a clear signal, $\sigma = 1/10$, for medium noise, $\sigma = 1/4$, and, for high noise, $\sigma = 1$. To guard for sampling effects, we repeat each procedure 50 times and average the results, which are displayed in Figure 1. One can see that, for the clear and medium signal, the algorithm can recover the differentially coexpressed genes modules reliably. Also, depending on the prominence of the signal, the influence of σ is more or less pronounced. In an exploratory analysis setting with several hidden patterns, we could use T_1, T_2, and T_3 to control the size of target patterns.

3.2. Analyzing a Type 2 Diabetes (T2D) in Rats. As a real-world application, we apply the BMKC method to a pair of type 2 diabetes (T2D) rats datasets (dataset pair T), which has been published in study [33]. Dataset pair T is from dataset GSE3068 of Gene Expression Omnibus (GEO) database. Yu et al. preprocessed dataset GSE3068. Dataset pair T includes 4765 genes in 10 disease samples and 10 normal samples. We use our algorithm to find differential coexpression modules in the type 2 diabetes.

For computational efficiency, we calculate the sum of each row or column of adjacency matrix; the sum means the number of genes related to the gene. The gene is outlier if the sum is zero. First, we calculate the sum of each row or column of the adjacency matrix and delete the outlier. Second, we calculate the sum of each row or column of the adjacency matrix and discard the lower 50% of them. We set $T_3 = 1.3$ and the minimum number of each clique to four. Finally, we apply our algorithm to the remainder genes and excavate two differential coexpression modules. Tables 1 and 2 list each gene symbol in the clique. The adjacency graphs of each differential coexpression module are shown in Figures 2 and 3. From these two figures, we can see that the cliques in each of the differential coexpression modules are overlapping, forming a closely related module. In normal condition, the absolute bicor value of total of 24 genes in modules distributes from 0.78 to 0.97. Yet, in disease condition, the absolute bicor value of genes distributes from 0.21 to 0.09. In the results of our study, the gene differential coexpression modules included quite a number of previously reported T2D-related genes: *Hifla* and *Sirt2* [34], *Smarca4* [35], *Sh2b2* [36], *Madd* [37], and *Rxrb* [38]. Despite not being previously reported to be related with T2D, other genes in the modules should receive adequate attention for their distinct traits from the perspective of differential coexpression. Further studies on the transcriptional mechanisms and functional consequences could pay more attention to these genes.

Table 1: Genes in each clique.

Clique number	Gene symbol				
1	***Hifla****	*Ifngr1*	*RGD1305094*	*Tenc1*	***Sirt2***
2	*Clcn1*	***Smarca4***	*Zkscan17*	*Rpl27a*	***Sirt2***
3	***Hifla***	*Ifngr1*	*Pfkfb3*	*Tenc1*	***Sirt2***
4	***Sh2b2***	*Pcsk5*	*Lamc1*	*Rpl27a*	***Sirt2***
5	*Lamc1*	***Smarca4***	*Zkscan17*	***Sirt2***	*Rpl27a*
6	***Hifla***	*RGD130504*	*Mxd4*	***Sirt2***	
7	*Tra1*	***Smarca4***	*Zkscan17*	***Sirt2***	

*Bold genes refer to the previously reported T2D-related genes. The other genes are identified in the differential coexpression modules.

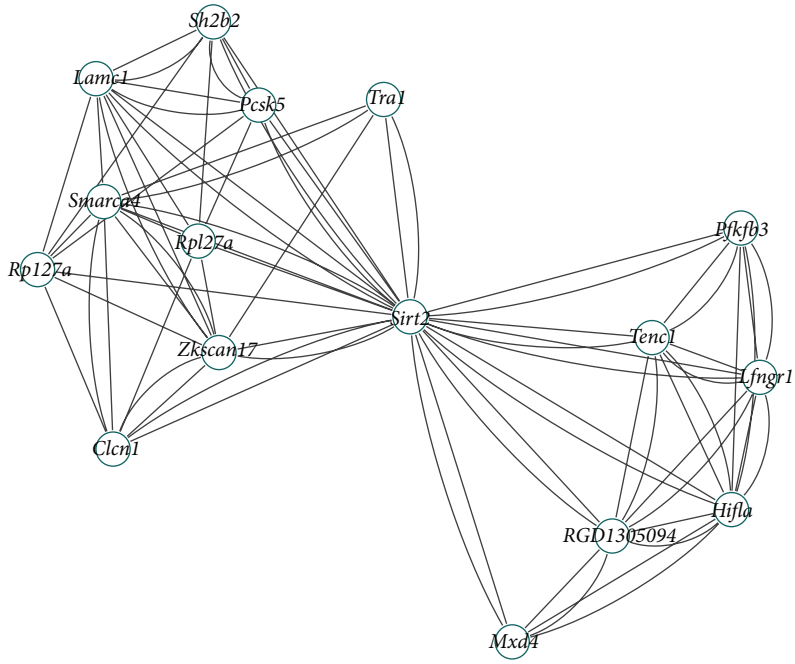

Figure 2: The adjacency graph of first gene differential coexpression module.

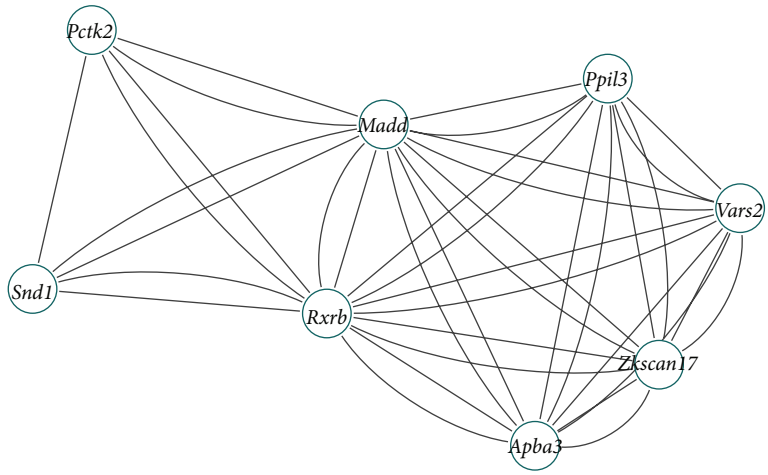

Figure 3: The adjacency graph of second gene differential coexpression module.

Table 2: Genes in each clique.

Clique number	Gene symbol					
1	*Vars2*	*Apba3*	***Madd***	*Zkscan17*	*Ppil3*	***Rxrb***
2	*Snd1*	***Madd***	*Pctk2*	***Rxrb***		

3.3. Significance Analysis of the BMKC Method. Naturally, the question of whether our findings are artifacts of the high dimensionality of the data arises. To assess this question, we apply a permutation procedure. Under the null hypothesis, we assume that all genes are mutually independent in both conditions groups. We heuristically sample from the null hypothesis by (group-wise) shuffling the expression values for each gene independently. Thus, random expression data are generated where all covariance structures are removed. Applying our algorithm to the randomized data yields one random score. We repeat the procedure 1000 times. Using the empirical distribution of the simulated scores, the simulated score means the global total sum of differential coexpression change of each gene in modules. We calculate P values for the observed scores in the nonpermuted data. For each of the patterns in the type 2 diabetes example, we only observe one random score smaller than the biological one. This corresponds to an empirical P value of 0.001. Hence, it is unlikely that the observed differential coexpression is a chance artifact.

4. Conclusions

In this paper, we proposed a new approach in gene sets level for differential coexpression analysis, which combine biweight midcorrelation and threshold selection strategy and also applied maximum clique concept with k-clique algorithm to the specific gene set to further investigate gene regulatory networks. Biweight midcorrelation is more robust for outliers and threshold selection strategy is an effective preprocess step of the proposed method. Experimental results on simulated datasets show that our method had good performance. We apply the proposed BMHT method to real dataset designed for T2D study, and two differential coexpression gene modules were detected, which should be a useful resource for T2D study and could be used for exploring the biological function of the related genes. In the future, we will focus on how to quickly excavate gene differential coexpression module from gene coexpression adjacency matrix.

Conflict of Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants nos. 61272339, 61271098, and 31301101 and the Key Project of Anhui Educational Committee, under Grant no. KJ2012A005.

References

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[2] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[3] M. P. S. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[4] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," *Bioinformatics*, vol. 18, no. 1, pp. 207–208, 2002.

[5] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, no. 24, pp. 4348–4355, 2005.

[6] J. Rachlin, D. D. Cohen, C. Cantor, and S. Kasif, "Biological context networks: a mosaic view of the interactome," *Molecular Systems Biology*, vol. 2, article 66, 2006.

[7] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpresion analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[8] A. Reverter, A. Ingham, S. A. Lehnert et al., "Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer," *Bioinformatics*, vol. 22, no. 19, pp. 2396–2404, 2006.

[9] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, "Gene co-expression network topology provides a framework for molecular characterization of cellular state," *Bioinformatics*, vol. 20, no. 14, pp. 2242–2250, 2004.

[10] M. J. Mason, G. Fan, K. Plath, Q. Zhou, and S. Horvath, "Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells," *BMC Genomics*, vol. 10, article 327, 2009.

[11] T. F. Fuller, A. Ghazalpour, J. E. Aten, T. A. Drake, A. J. Lusis, and S. Horvath, "Weighted gene coexpression network analysis strategies applied to mouse weight," *Mammalian Genome*, vol. 18, no. 6-7, pp. 463–472, 2007.

[12] J. M. Freudenberg, S. Sivaganesan, M. Wagner, and M. Medvedovic, "A semi-parametric Bayesian model for unsupervised differential co-expression analysis," *BMC Bioinformatics*, vol. 11, article 234, 2010.

[13] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego, Calif, USA, 1997.

[14] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 5, pp. 415–426, 2000.

[15] D. Kostka and R. Spang, "Finding disease specific alterations in the co-expression of genes," *Bioinformatics*, vol. 20, supplement 1, pp. i194–i199, 2004.

[16] C. Prieto, M. J. Rivas, J. M. Sánchez, J. López-Fidalgo, and J. de Las Rivas, "Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes," *Bioinformatics*, vol. 22, no. 9, pp. 1103–1110, 2006.

[17] I. C. Ross and F. Harary, "On the determination of redundancies in sociometric chains," *Psychometrika*, vol. 17, no. 2, pp. 195–208, 1952.

[18] S. Wasserman and K. Faust, *Social Network Analysis, Methods and Applications*, Cambridge University Press, Cambridge,UK, 1994.

[19] Z. Qibing, H. Yuan, Z. Qimin, S. Yan, and L. Zhihua, "Role for Krüppel-like factor 4 in determining the outcome of p53 response to DNA damage," *Cancer Research*, vol. 69, no. 21, pp. 8284–8292, 2009.

[20] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of Combinatorial Optimization*, pp. 1–74, 1999.

[21] B. Albert-László and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] J. Scott, *Social Network Analysis: A Handbook*, Sage Publications, London, UK, 2nd edition, 2000.

[23] R. M. Shiffrin and K. Borner, "Mapping knowledge domains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, supplement 1, pp. 5183–5185, 2004.

[24] B. S. Everitt, *Cluster Analysis*, Edward Arnold, London, UK, 3th edition, 1993.

[25] S. Knudsen, *A Guide to Analysis of DNA Microarray Data*, Wiley-Liss, New York, NY, USA, 2nd edition, 2004.

[26] M. E. J. Newman, "Detecting community structure in networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.

[27] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[28] M. G. Everett and S. P. Borgatti, "Analyzing clique overlap," *Connections*, vol. 21, pp. 49–61, 1998.

[29] S. Kosub, "Local density," in *Network Analysis*, pp. 112–142, Springer, Berlin, Germany, 2005.

[30] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[31] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, 5 pages, 2004.

[32] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letters*, vol. 94, no. 16, Article ID 160202, 2005.

[33] H. Yu, B.-H. Liu, Z.-Q. Ye, C. Li, Y.-X. Li, and Y.-Y. Li, "Link-based quantitative methods to identify differentially coexpressed genes and gene Pairs," *BMC Bioinformatics*, vol. 12, article 315, 2011.

[34] J. C. Milne, P. D. Lambert, S. Schenk et al., "Small molecule activators of SIRT1 as therapeutics for the treatment of type 2 diabetes," *Nature*, vol. 450, no. 7170, pp. 712–716, 2007.

[35] L. L. Nguyen, A. D. Kriketos, D. P. Hancock, I. D. Caterson, and G. S. Denyer, "Insulin resistance does not influence gene expression in skeletal muscle," *Journal of Biochemistry and Molecular Biology*, vol. 39, no. 4, pp. 457–463, 2006.

[36] M. Li, Z. Li, D. L. Morris, and L. Rui, "Identification of SH2B2β as an inhibitor for SH2B1- and SH2B2α-promoted Janus kinase-2 activation and insulin signaling," *Endocrinology*, vol. 148, no. 4, pp. 1615–1621, 2007.

[37] J. Dupuis, C. Langenberg, I. Prokopenko, and R. Saxena, "The genetics of type 2 diabetes: what have we learned from GWAS?" *Nature*, vol. 1212, pp. 59–77, 2010.

[38] S. Sookoian and C. J. Pirola, "Metabolic syndrome: from the genetics to the pathophysiology," *Current Hypertension Reports*, vol. 13, no. 2, pp. 149–157, 2011.