BMC Bioinformatics

RESEARCH Open Access

An effective method for network module extraction from microarray data

Priyakshi Mahanta[1], Hasin A Ahmed[1], Dhruba K Bhattacharyya[1*], Jugal K Kalita[2]

From The 8th Annual Biotechnology and Bioinformatics Symposium (BIOT-2011)
Houston, TX, USA. 20-21 October 2011

Abstract

Background: The development of high-throughput Microarray technologies has provided various opportunities to systematically characterize diverse types of computational biological networks. Co-expression network have become popular in the analysis of microarray data, such as for detecting functional gene modules.

Results: This paper presents a method to build a co-expression network (CEN) and to detect network modules from the built network. We use an effective gene expression similarity measure called NMRS (Normalized mean residue similarity) to construct the CEN. We have tested our method on five publicly available benchmark microarray datasets. The network modules extracted by our algorithm have been biologically validated in terms of Q value and p value.

Conclusions: Our results show that the technique is capable of detecting biologically significant network modules from the co-expression network. Biologist can use this technique to find groups of genes with similar functionality based on their expression information.

Introduction

The development of high-throughput Microarray technologies has provided a range of opportunities to systematically characterize diverse types of biological networks. Biological networks can be broadly classified as protein interaction networks [1-3], metabolic networks [4-6] and gene co-expression networks [7]. These networks provide an effective way to summarize gene and protein correlations. In this paper, we focus on gene co-expression networks, which is an undirected graph where nodes represent gene and nodes are connected by an edge if the corresponding gene pairs are significantly co-expressed. Gene co-expression networks provide the association between individual genes in terms of their expression similarity and a network-level view of the similarity among a set of genes. In co-expression networks, two genes are connected by an undirected edge if their activities have significant association, as computed using gene expression measurements such as Pearson correlation, Spearman correlation, mutual information. Compared to gene regulatory networks, a gene co-expression network is built upon gene neighborhood relations, which give interesting geometric interpretations of the network. One of the most important applications of gene co-expression networks is to identify functional gene modules [8] or network modules, which are represented by the strongly connected regions of the co-expression network.

Problem formulation

Due to non-transitive nature of connections among genes, genes form a very complicated connectivity network with respect to a particular similarity measure in a gene expression data set. Such a connectivity network is often referred to as a co-expression network. A major use of this co-expression network is extraction of network modules that represent the strongly connected regions in the co-expression network. These modules may present highly co expressed genes, which are functionally similar.

In this paper, we propose an effective similarity measure for gene co-expression, develop an approach to prepare a co- expression network from a gene expression

* Correspondence: dkb@tezu.ernet.in
[1]Dept. of Comp. Sc. and Engg, Tezpur University, Napaam, Tezpur, India
Full list of author information is available at the end of the article

data set and mine the potential network modules from the built network. We aim to produce a graph, $G=\{V,E\}$ that presents the co-expression network with the following properties.

1. Each vertex $v\in V$ represents a gene.

2. Each edge $e\in E$ represents a connection between a pair of vertices v_1,v_2 where $v_1,v_2 \in V$.

3. There is an edge between two vertices $v_1,v_2 \in V$ if the similarity of the genes corresponding to the vertices is more than a user defined threshold.

Our contribution

We claim the following contributions in this paper.

- We introduce an effective gene similarity measure NMRS.
- We propose an approach to construct a co-expression network using NMRS.
- We develop a spanning tree based method to extract the potential network modules.

Background

In the literature, a number of techniques have been proposed for gene co-expression network construction. When inferring co-expression networks from gene expression data, the algorithms take a gene expression dataset as primary input and then, by using a correlation-based proximity measure, constructs the corresponding co-expression networks. Frequently used correlation-based measures are Pearson correlation coefficient, Spearman correlation coefficient and Mutual information. Approaches such as [9,10] used Pearson correlation coefficient to extract the association among genes in a co-expression network. The Spearman correlation coefficient is used as a gene expression similarity measure to construct co-expression network in [10]. [11], Steuer et al. [12] reports the use of Mutual Information to find similarly expressed gene pairs in such networks. While some studies attempted to apply algorithms directly to the adjacency matrices of networks to partition network nodes into groups [13,14], other studies rely on special purpose algorithms for identifying subnetworks with certain properties [15].

Generally, in a co-expression network, the connections between genes are obtained from the absolute values of a co-expression measure. Several researchers have suggested to threshold this value of the co-expression measure to construct gene co-expression networks. There are two ways to pick a threshold: one way is picking a hard threshold (a number) based on the notion of statistical significance so that gene co-expression is encoded using binary information (connected=1, unconnected=0). The other way is called soft thresholding which weighs each connection by a number between 0 and 1. The drawbacks of hard thresholding include loss of information regarding the magnitude of gene connections and sensitivity to the choice of the threshold. Generally, hard thresholding results in unweighted networks while soft thresholding results in weighted networks.

Methodology

To construct the gene co-expression network, we use the general framework proposed by [16]. A new effective gene similarity measure called NMRS is used to construct the distance matrix. We use a hard thresholding based signum function to construct the adjacency matrix from the distance matrix. A spanning tree based approach is used to detect network modules in the co-expression network. Extracted network modules are projected as functional categories of genes and these modules are validated using p value and Q value. Our approach is explained next.

Define a gene expression measurement

To determine whether two genes have similar expression patterns, an appropriate similarity measure must be chosen [17]. To measure the level of concordance between gene expression profiles, we develop a gene co-expression measure called NMRS. The NMRS of gene $d_1=(a_1, a_2,..., a_n)$ with respect to gene $d_2=(b_1, b_2,..., b_n)$ is defined by

$$\mathrm{NMRS}(d_1,d_2)=1-\frac{\sum_{i=1}^{p}|a_i-a_{mean}-b_i+b_{mean}|}{2\times max\left\{\sum_{i=1}^{p}|(a_i-a_{mean})|,\sum_{i=1}^{p}|(b_i-b_{mean})|\right\}}$$

where

$$a_{mean} \text{ is the mean of all the elements of gene } d_1;$$
$$a_{mean}=\{a_1+a_2+\ldots+a_n\}/n,$$
$$b_{mean} \text{ is the mean of all the elements of gene } d_2 and$$
$$b_{mean}=\{b_1+b_2+\ldots+b_n\}/n.$$

NMRS as a metric

NMRS satisfies all the properties of a metric. We establish The non-negativity, symmetricity and triangular inequality properties for our measure in Additional file 1.

Significance of NMRS

The most widely used proximity measures in gene expression data analysis are Euclidean distance, Pearson correlation coefficient, Spearman correlation coefficient, Mean squared residue etc. In co-expression network, the used proximity measure is expected to effectively detect the linear shifting patterns in the gene expression data. But none of the widely used proximity measures can satisfactorily serve this purpose. The Euclidean distance measures the distance between two data objects. But in this domain, the overall shapes of gene expression patterns (or profiles) are of greater interest than the individual magnitudes of each feature [18]. So Euclidean distance can not straight away detect shifting patterns, but bringing down all the genes to the same range of expression values can make this measure to detect shifting patterns. This normalization process involves an extra overhead. Along with shifting patterns Pearson correlation coefficient also detects scaling patterns and some other patterns which is normally not desired in a co-expression network and may lead to inclusion of genes which have considerable amount of difference between their expression levels. Spearman Rank Correlation Coefficient uses ranks to calculate correlation which can neither detect shifting patterns nor scaling patterns. Mean squared residue is good enough to detect shifting patterns, but the aggregate measure can not operate in a mutual mode, i.e. it can not find correlation between a pair of genes. A general comparison of these measures is presented in Table 1.

Let us consider a random gene pattern *a* as presented in Figure 1(a). Gene pattern *b1* in Figure 1(b) has a shifting relationship with gene *a.* Gene pattern *b8* in Figure 1(i) is a shifted as well as negatively correlated form of gene *a.* Figures 1(b)-1(h) present gene patterns *b2*, *b3*, *b4*, *b5*, *b6* and *b7* which are uniformly distributed intermediate patterns between genes *b1* and *b8.* Figure 2 shows Pearson, Spearman and NMRS correlation of gene patterns *b1-b8* with that of *a.* As usual the Spearman correlation was found to be concerned only about the rank information about the gene patterns. Interestingly, Pearson correlation was found to produce some undesired correlation values for the pairs *a and b2*, *a and b3*, *a and b4*, *a and b4*, *a and b5*, *a and b6* and *a and b7*, which are neither shifting nor scaling patterns. The values of these patterns are given in Table 2. Our measure is found to effectively distinguish patterns across this uniform distribution from a shifted pattern (with a value 1) to a shifted and negatively correlated pattern (with value 0) of a given pattern as can be seen in Figure 2.

Compute an adjacency matrix

An adjacency matrix is obtained using a signum function based hard thresholding approach which encodes edge information for each pair of nodes in the co-expression network. Two genes d_i and d_j are connected if Dist(d_i,d_j) >δ, a user defined threshold. Based on the connected pairs, an adjacency matrix is computed as

$$A(i,j) = \begin{cases} 1 & \text{if } d_i \text{ and } d_j \text{ are connected;} \\ 0 & \text{otherwise.} \end{cases}$$

Detect network modules

To detect subsets of nodes (modules) that are tightly connected to each other is an important aim of co-expression network analysis. In this paper, we use spanning trees and a topological overlap similarity measure [19] to find the network modules, since this measure is found to result in biologically meaningful modules. A tree *T* is a spanning tree of a connected graph *G* if *T* is a subgraph of *G* and it contains all vertices of *G.* We use Prim's algorithm [20] to find a spanning tree of a undirected graph. However, unlike traditional Prim's algorithm we find a spanning tree with maximum weight. For unweighed networks (i.e. a_{ij} = 1 or = 0), the topological overlap matrix is defined by

$$w_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{1}$$

where $l_{ij} = \Sigma_u a_{iu} a_{ij}$, and $k_i = \Sigma_u a_{iu}$ is the node connectivity.

Extract useful information

Extraction of useful biological information is one of the main usages of gene co-expression networks. From the constructed network, one can explore various important information such as functionality and pathways of genes, essential genes susceptible to diseases.

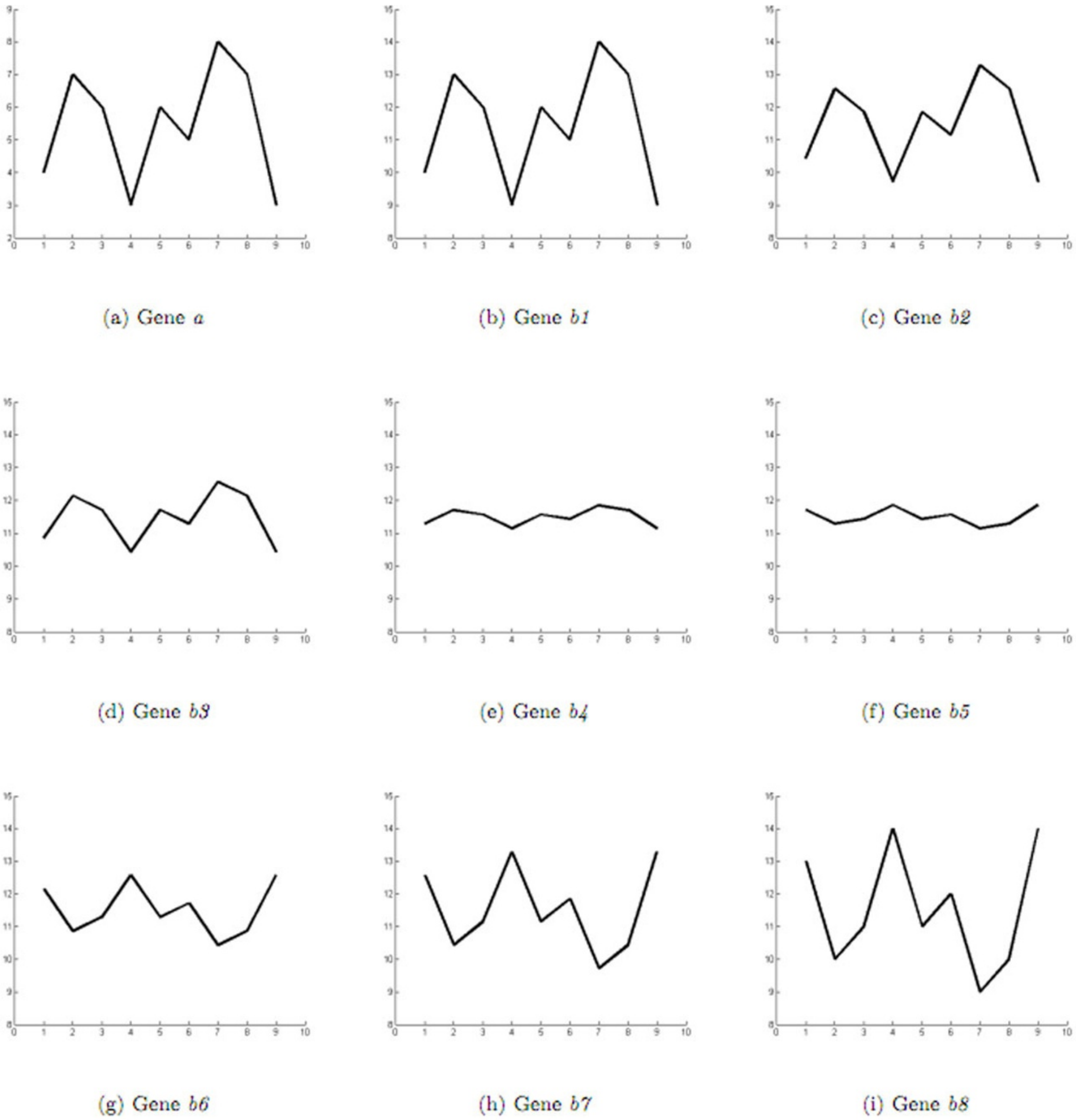
Proposed algorithm: Module Miner

Module Miner takes NMRS threshold, δ, as a input and works on a microarray gene data and constructs the gene co-expression network and finally network modules are extracted from the network. Our approach uses an effective similarity measure NMRS to form a co-expression network using signum function. The co-expression network is further explored to mine the

Table 1 Comparison of proximity measures

Proximity measure	*Mode*	*Normalization required*	*Detects shifting pattern*	*Detects scaling pattern*
Euclidian	Mutual	Yes	Yes	No
Pearson	Mutual	No	Yes	Yes
Spearman	Mutual	No	No	No
MSR	Aggregate	No	Yes	Yes
NMRS	Mutual	No	Yes	Yes

The table 1 presents the comparison of different proximity measure.



(a) Gene *a*

(b) Gene *b1*

(c) Gene *b2*

(d) Gene *b3*

(e) Gene *b4*

(f) Gene *b5*

(g) Gene *b6*

(h) Gene *b7*

(i) Gene *b8*

Figure 1: gene *b1* is shifted form of gene a. gene b8 is negatively correlated and shifted form of gene *a.* gene *b2, b3, b4, b5, b6* and *b7* are genes having uniformly distributed intermediate patterns between gene *b1* and *b8.*

Figure 1 Example patterns used for evaluation of proximity measures The figure 1 presents the value of some example patterns that are used to demonstrate the superiority NMRS over other proximity measures viz. Euclidean distance, Pearson correlation coefficient and Spearman correlation coefficient.

potential network modules using a spanning tree based method and a connectivity measure called *Topological Overlap Matrix.*

The symbols provided in Table 3 and definitions given below are useful in discussing the proposed Module Miner algorithm.

Definition 1 *A* ***CEN*** *can be defined by an undirected, graph* $G=\{V,E\}$ *where each* $v\in V$ *corresponds to a gene and each edge* $e\in E$ *corresponds a pair of genes* d_i, $d_j\in D$ *such that* $Dist(d_{\mathrm{i}}, d_j)\geq\delta$.

Definition 2 ***Connected regions*** *in a CEN are parts of the network where each pair of vertices is connected*

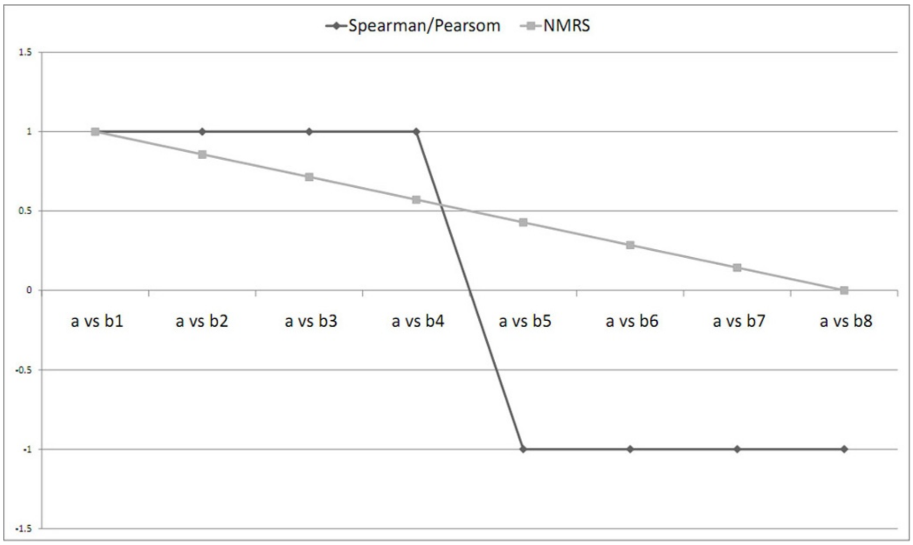


Figure 2 NMRS and Pearson correlation coefficient among considered example patterns The figure 2 presents NMRS and Pearson correlation coefficient of patterns *b1-b8* with that of *a*.

by a path. The i^{th} connected region extracted from G can be defined as a graph $G_i^{con} = \{V_i^{con}, E_i^{con}\}$ where $E_i^{con} \subseteq E$ and $E_i^{con} \subseteq E$ such that for any vertex $v_i \in V_i^{con}$, there is at least one vertex $v_j \in V_i^{con}$ which are connected by an edge $e \in E_i^{con}$.

Definition 3 ***Maximum spanning tree*** *G_i^{span} of a weighted graph is a spanning tree obtained from i^{th} connected region, G_i^{span} can be defined as $\{V_i^{con}, E_i^{span}\}$, where the sum of TOM values associated with edges in E_i^{span} is maximum compared to other spanning trees.*

Definition 4 ***Network modules*** *are highly connected regions of the co-expression network. The i^{th} network module derived from j^{th} connected region G_j^{con} is defined as a set of vertices V_i^{net} if*

- *$TOM(V_1) > TOM(V_i^{net})$ and $V_1, V_2 \subseteq V_i^{net}$ where $V_1, V_2 \subseteq V_i^{net}$ are obtained by removing the weakest edge of the maximum spanning tree built for the subgraph of G consisting of vertex set V_i^{net} or*
- *$TOM(V_3) > TOM(V_4)$ and $V_3, V_i^{net} \subseteq V_4$ where, $V_3, V_i^{net} \subseteq V_4$ are obtained by removing the weakest edge of the maximum spanning tree built for the subgraph of G consisting of vertex set V_4.*

Algorithm: *Module Miner*

The pseudo code of Module Miner is presented in Algorithm 1. In the pseudo code, lines 1-4 extracts the connected regions from the gene expression data. Lines 5-25 process each of the connected regions to extract the network modules. A maximum spanning tree is constructed using Prim's algorithm [20] from a connected region with weights defined by topological overlap matrix in lines 6-8. Lines 9-10 find and remove the weakest edge from the spanning tree. Removal of this edge from the spanning tree leads to two subtrees which are processed in lines 11-23 to form either a connected module or a new connected region.

Table 2 Gene pattern

a	4	7	6	3	6	5	8	7	3
b1	10	13	12	9	12	11	14	13	9
b2	10.4286	12.5714	11.8571	9.7143	11.8571	11.1429	13.2857	12.5714	9.7143
b3	10.8571	12.1429	11.7143	10.4286	11.7143	11.2857	12.5714	12.1429	10.4286
b4	11.2857	11.7143	11.5714	11.1429	11.5714	11.4286	11.8571	11.7143	11.1429
b5	11.7143	11.2857	11.4286	11.8571	11.4286	11.5714	11.1429	11.2857	11.8571
b6	12.1429	10.8571	11.2857	12.5714	11.2857	11.7143	10.4286	10.8571	12.5714
b7	12.5714	10.4286	11.1429	13.2857	11.1429	11.8571	9.7143	10.4286	13.2857
b8	13	10	11	14	11	12	9	10	14

The table 2 presents the random gene patterns for analysis of different proximity measures.

Table 3 Symbolic representation

SYMBOL	*MEANING*
D	The gene expression matrix
d_i	ithgene in D
δ	Signum threshold
G	Co-expression network
V	Set of vertices in G
E	Set of edges in G
Dist	Distance matrix
$Dist(d_i, d_j)$	NMRS distance between genes d_i, $d_j \in$ D
Adj	Adjacency matrix
$Adj(v_i,v_j)$	1 if v_i and v_j are connected by an edge 0 otherwise
G^{con}	Set of connected region
G_i^{con}	ith connected region
V_i^{con}	Set of vertices in ith connected region
E_i^{con}	Set of edges in ith connected region
Adj_i^{con}	Adjacency matrix of the ith connected region
G_i^{net}	ith network module
D^{net}	Set of network modules obtained from *G*
$TOM(v_i,v_j)$	Topological Matrix value between vertices v_i and v_j
$TOM(V_1)$	Average TOM of the set of vertices V_1
TOM_i^{con}	TOM for ith connected region
G_i^{span}	Maximum spanning tree obtained from ith connected region
E_i^{span}	Set of edges in G_i^{span}

The table 3 describes the various symbols that is used in *ModuleMiner*.

```
Input : D and $\delta$
Output : $G$, $D^{net}$

Prepare $Dist$ using NMRS;
Prepare $Adj$ from $Dist$ using signum function with threshold $\delta$;
Build G from $Adj$;
Find $G^{con}$ from G;
foreach $G_i^{con} \in G^{con}$ do
    Prepare $Adj_i^{con}$;
    Prepare $TOM_i^{con}$ from $Adj_i^{con}$;
    Mine $G_i^{span}$ using $TOM_i^{con}$;
    Find $e \in E_i^{span}$ with smallest TOM value;
    Remove $e$;
    If $G_1 = \{V_1, E_1\}$, $G_2 = \{V_2, E_2\} \subset G_i^{span}$ are the subtrees obtained after removing $e$, process $G_1$ and
    $G_2$ as followed;
    if $TOM(V_1) > TOM(V_i^{con})$ then
        Add $G_1$ to $G^{con}$;
        if $TOM(V_2) > TOM(V_i^{con})$ then
            Add $G_2$ to $G^{con}$
        end
        else
            Add genes corresponding to vertices in $V_2$ to $D^{net}$ as a new network module
        end
    end
    else
        if $TOM(V_2) \leq TOM(V_i^{con})$ then
            Add genes corresponding to vertices $V_i^{con}$ to $D^{net}$ as new network module
        end
        else
            Add genes corresponding to vertices in $V_1$ to $D^{net}$ as a new network module;
            Add $G_1$ to $G^{con}$
        end
    end
end
```

Algorithm 1 : Module Miner

Algorithm complexity

The complexity of different steps of our method is presented in this section.

- The preparation of the distance matrix involves a complexity of $O(n \times n\text{-}1)/2$, where n is the number of genes.
- Finding connected regions from the co-expression network requires a complexity of $O(n)$.
- Computation of the TOM matrix involves a complexity of $O(n_c \times (d_c \times (d_c\text{-}1)/2))$, where n_c is the total number of connected regions and d_c is the average number of genes in the connected regions.
- Finding a maximum spanning tree consumes a complexity of $O(n_c \times d_c^2)$.

Experimental results

We implemented the Module Miner algorithm in MATLAB and tested it on five benchmark microarray datasets mentioned in Table 4. The test platform was a SUN workstation with Intel(R) Xenon(R) 3.33 GHz processor and 6 GB memory running Windows XP operating system.

Validation

The performance of Module Miner on the five publicly available benchmark microarray dataset is measured in terms of p value and Q value.

p value

Biological significance of the sets of genes included in the extracted network modules are evaluated based on p values [21]. p value signifies how well these genes match with different Gene Ontology(GO) categories. A cumulative hypergeometric distribution is used to compute the p value. A low p-value of the set of genes in a network module indicates that the genes belong to enriched functional categories and are biologically significant. From a given GO category, the probability p of getting k or more genes within a cluster of size n, is defined as

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} \tag{2}$$

where f and g denote the total number of genes within a category and within the genome respectively.

To compute p-value, we used a tool called FuncAssociate [22]. FuncAssociate computes the hyper geometric functional enrichment score based on Molecular Function and Biological Process annotations. The enriched functional categories for some of the network modules obtained by Module miner on the datasets are presented

Table 4 Datasets used for evaluating *ModuleMiner*

Serial. No	Dataset	No. of Genes/ No. of Conditions	Source
1	Yeast Sporulation	474/17	http://cmgm.stanford.edu/pbrown/sporulation/index.html
2	Yeast Diauxic Shift	689/72	Sample gene in expander
3	Subset of Yeast Cell Cycle	384/17	http://faculty.washington.edu/kayee/cluster
4	Arabidopsis Thaliana	138/8	http://homes.esat.kuleuven.be/~sistawww/bioi/thijs/Work/Clustering.html
5	Rat CNS	112/9	http://faculty.washington.edu/kayee/cluster

The table 4 gives the description of various datasets used in *ModuleMiner*.

in Tables 5 and 6. The co-expression network modules produced by Module Miner contains the highly enriched cellular components of *DNA replication, DNA repair, DNA metabolic process, response to DNA damage stimulus, nuclear nucleosome, nucleosome, nucleosome assembly, protein-DNA complex, cell wall assembly, meiosis, cell differentiation, sporulation resulting in formation of a cellular spore, sporulation, anatomical structure formation involved in morphogenesis, cellular developmental process, reproductive cellular process, cell cycle phase, developmental process, cell cycle process*etc with p-values of $\mathbf{7.69 \times 10^{-27}}$, $\mathbf{3.93 \times 10^{-25}}$, $\mathbf{1.03 \times 10^{-26}}$, $\mathbf{1.23 \times 10^{-23}}$, $\mathbf{2.32 \times 10^{-28}}$, $\mathbf{5\ .12 \times 10^{-27}}$, $\mathbf{7.27 \times 10^{-23}}$, $\mathbf{2.06 \times 10^{-20}}$, $\mathbf{3.84 \times 10^{-32}}$, $\mathbf{1.41 \times 10^{-31}}$, $\mathbf{1.19 \times 10^{-38}}$, $\mathbf{9.65 \times 10^{-36}}$, $\mathbf{1.34 \times 10^{-20}}$, $\mathbf{2.52 \times 19^{-34}}$, $\mathbf{1.93 \times 10^{-28}}$ **and** $\mathbf{6.91 \times 10^{-27}}$ being the highly enriched one. From the given p values, we can conclude that Module Miner shows a good enrichment of functional categories and therefore project a good biological significance.

Table 5 P-value of one of the network modules of Dataset 2

P-value	*GO number*	*GO category*
2.32E-28	GO:0000788	nuclear nucleosome
5.12E-27	GO:0000786	nucleosome
7.27E-23	GO:0006334	nucleosome assembly
2.06E-20	GO:0032993	protein-DNA complex
8.61E-19	GO:0034728	nucleosome organization
1.14E-18	GO:0065004	protein-DNA complex assembly
1.12E-17	GO:0006333	chromatin assembly or disassembly
4.12E-16	GO:0005694	chromosome
2.49E-14	GO:0044454	nuclear chromosome part
1.70E-13	GO:0031298	replication fork protection complex
9.47E-14	GO:0006325	chromatin organization
6.78E-13	GO:0044427	chromosomal part
2.32E-12	GO:0034622	cellular macromolecular complex assembly

The table 5 gives the p value of one of the network modules of Dataset 2.

Table 6 p-value of one of the network modules of Dataset 3

P-value	*GO number*	*GO category*
3.93E-25	GO:0006281	DNA repair
1.03E-26	GO:0006259	DNA metabolic process
1.23E-23	GO:0006974	response to DNA damage stimulus
7.69E-27	GO:0006260	DNA replication
6.94E-19	GO:0007049	cell cycle
5.55E-16	GO:0005634	nucleus
8.53E-18	GO:0044454	nuclear chromosome part
1.51E-17	GO:0022402	cell cycle process
3.53E-17	GO:0000079	regulation of cyclin-dependent protein kinase activity
5.72E-15	GO:0045859	regulation of protein kinase activity
5.16E-16	GO:0005657	replication fork

The table 6 gives the p value of one of the network modules of Dataset 3.

Q value

The Q-value [23] for a particular gene G is the proportion of false positives among all genes that are as or more extremely differentially expressed. Equivalently, the Q-value is the minimal False Discovery Rate(FDR) at which this gene appears significant. The GO categories and Q-values from a FDR corrected hypergeometric test for enrichment are reported in GeneMANIA. Q-values are estimated using the Benjamini Hochberg procedure. Different GO categories of the co-expression networks produced by Module miner are displayed up to a Q-value cutoff of 0.1 in Table 7, 8, 9, 10 and 11. The co-expression network modules produced by Module Miner contains the highly enriched cellular components of *sporulation resulting in formation of a cellular spore, spore wall assembly, ascospore wall assembly, ascospore formation, sexual sporulation, spore wall biogenesis, ascospore wall biogenesis, sexual sporulation resulting in formation of a cellular*

Table 7 Q-value of one of the network modules of Dataset 3

GO annotation	*Q value*
DNA replication	1.93E-21
DNA repair	1.93E-21
response to DNA damage stimulus	2.17E-20
DNA-dependent DNA replication	3.07E-19
replication fork	6.27E-19
nuclear chromosome	1.23E-17
mitotic sister chromatid cohesion	5.51E-17
nuclear replication fork	9.37E-17
nuclear chromosome part	2.00E-16
sister chromatid cohesion	5.13E-15

The table 7 gives the Q value of one of the network modules of Dataset 3.

Table 8 Q-value of one of the network modules of Dataset 1

GO annotation	*Q value*
cytosolic ribosome	1.43E-52
cytosolic part	3.26E-48
structural constituent of ribosome	2.11E-44
ribosomal subunit	1.16E-42
cytosolic large ribosomal subunit	2.65E-36
large ribosomal subunit	1.47E-27
preribosome	2.96E-23
cytosolic small ribosomal subunit	3.71E-17
90S preribosome	8.48E-16

The table 8 gives the Q value of one of the network modules of Dataset 1.

*spore, cell development cell wall assembly, reproductive process in single-celled organism, cell differentiation, fungal-type cell wall biogenesis, reproductive developmental process, reproductive process, reproductive cellular process, reproduction of a single-celled organism, cell wall biogenesis, sexual reproduction, anatomical structure development, anatomical structure morphogenesis , M phase, meiotic cell cycle, meiosis, M phase of meiotic cell cycle*etc with Q-values of $\mathbf{1.53 \times 10^{-34}}$, $\mathbf{3.43 \times 10^{-33}}$, $\mathbf{2.59 \times 10^{-32}}$, $\mathbf{6.93 \times 10^{-30}}$, $\mathbf{1.40 \times 10^{-29}}$, $\mathbf{1.86 \times 10^{-25}}$, $\mathbf{9.90 \times 10^{-25}}$, $\mathbf{1.25 \times 10^{-24}}$, $\mathbf{4.83 \times 10^{-24}}$, $\mathbf{5.45 \times 10^{-24}}$, $\mathbf{2.10 \times 10^{-23}}$, $\mathbf{1.62 \times 10^{-21}}$, $\mathbf{2.74 \times 10^{-21}}$ being the highly enriched one. From the results of Q value, we arrive at the conclusion that the genes in a network module cluster obtained by Module Miner seem to be involved in similar functions.

Table 9 Q-value of one of the network modules of Dataset 1

GO annotation	*Q value*
sporulation resulting in formation of a cellular spore	1.53E-34
sporulation	1.53E-34
anatomical structure formation involved in morphogenesis	1.53E-34
spore wall assembly	3.43E-33
ascospore wall assembly	3.43E-33
ascospore formation	3.43E-33
sexual sporulation	3.43E-33
spore wall biogenesis	3.43E-33
ascospore wall biogenesis	3.43E-33
sexual sporulation resulting in formation of a cellular spore	3.43E-33
cell development	3.43E-33
cell wall assembly	8.88E-33
reproductive process in single-celled organism	2.59E-32
cell differentiation	8.40E-32
fungal-type cell wall biogenesis	6.93E-30
reproductive developmental process	1.40E-29
reproductive process	1.86E-25
reproductive cellular process	1.86E-25
reproduction of a single-celled organism	9.90E-25
cell wall biogenesis	1.25E-24
sexual reproduction	4.83E-24
anatomical structure development	5.45E-24
anatomical structure morphogenesis	5.45E-24
M phase	2.10E-23
meiotic cell cycle	1.62E-21
meiosis	2.74E-21
M phase of meiotic cell cycle	2.74E-21

The table 9 gives the Q value of one of the network modules of Dataset 1.

Table 10 Q-value of one of the network modules of Dataset 4

GO annotation	*Q value*
synaptic transmission	1.29E-13
glutamate receptor activity	3.77E-11
synapse	6.68E-08
regulation of synaptic transmission	3.06E-07
regulation of transmission of nerve impulse	4.00E-07
regulation of neurological system process	7.07E-07
regulation of system process	5.38E-05
synapse part	8.11E-04
cell projection part	9.46E-04

The table 10 gives the Q value of one of the network modules of Dataset 4.

We have used GeneMANIA [24] which is a flexible, user-friendly web interface for generating hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays. Given a query list, GeneMANIA extends the list with functionally similar genes that it identifies using available genomics and proteomics data. GeneMANIA displays results as an interactive network, illustrating the functional relatedness of the query and retrieved genes. GeneMANIA currently supports different networks including co-expression, physical interaction, genetic interaction, co-localization etc. On a given set of genes and their connectivity information, GeneMANIA also assigns coverage ratios as percentage to each of these networks with respect to the annotated

Table 11 Q-value of one of the network modules of Dataset 5

GO annotation	*Q value*
regulation of synaptic transmission	6.438756E-7
regulation of transmission of nerve impulse	9.297736E-7
regulation of neurological system process	1.533111E-6
intermediate filament cytoskeleton organization	2.056912E-6
intermediate filament-based process	5.218967E-6
neurofilament cytoskeleton	1.109702E-5
intermediate filament organization	1.454524E-5
synapse part	2.543099E-5
growth factor binding	2.571707E-5
intermediate filament	2.938762E-5
positive regulation of neurogenesis	9.6019E-5

The table 11 gives the Q value of one of the network modules of Dataset 5.

Table 12 The weightage of co-expression by Module Miner

Datasets	*Network Modules*	*Percentage*
Dataset1	C1	99.57%
	C2	88.89%
Dataset2	C1	59.23%
	C2	77.27%
Dataset3	C1	92.13%
	C2	88.89%
	C3	92.33%
	C4	67.65%
Dataset4	C1	81.85%
Dataset5	C1	76.62%

The table 12 gives the percentage of co-expression on network modules produced by Module Miner.

genes in the genome. The percentage of co-expression on network modules produced by Module Miner is given in Table 12. The values are obtained by choosing the default network weighting option i.e. **automatically selected weighing method**. Visualization of some of the co-expression networks generated by GeneMANIA for the datasets are presented in Figures 3, 4, 5.

Conclusion and future work

In this paper, an effective gene expression similarity measure NMRS is introduced, which is used to construct the co-expression network through a signum function based hard thresholding scheme. Finally, network modules are extracted from the network using maximum spanning tree and topological overlap matrix.

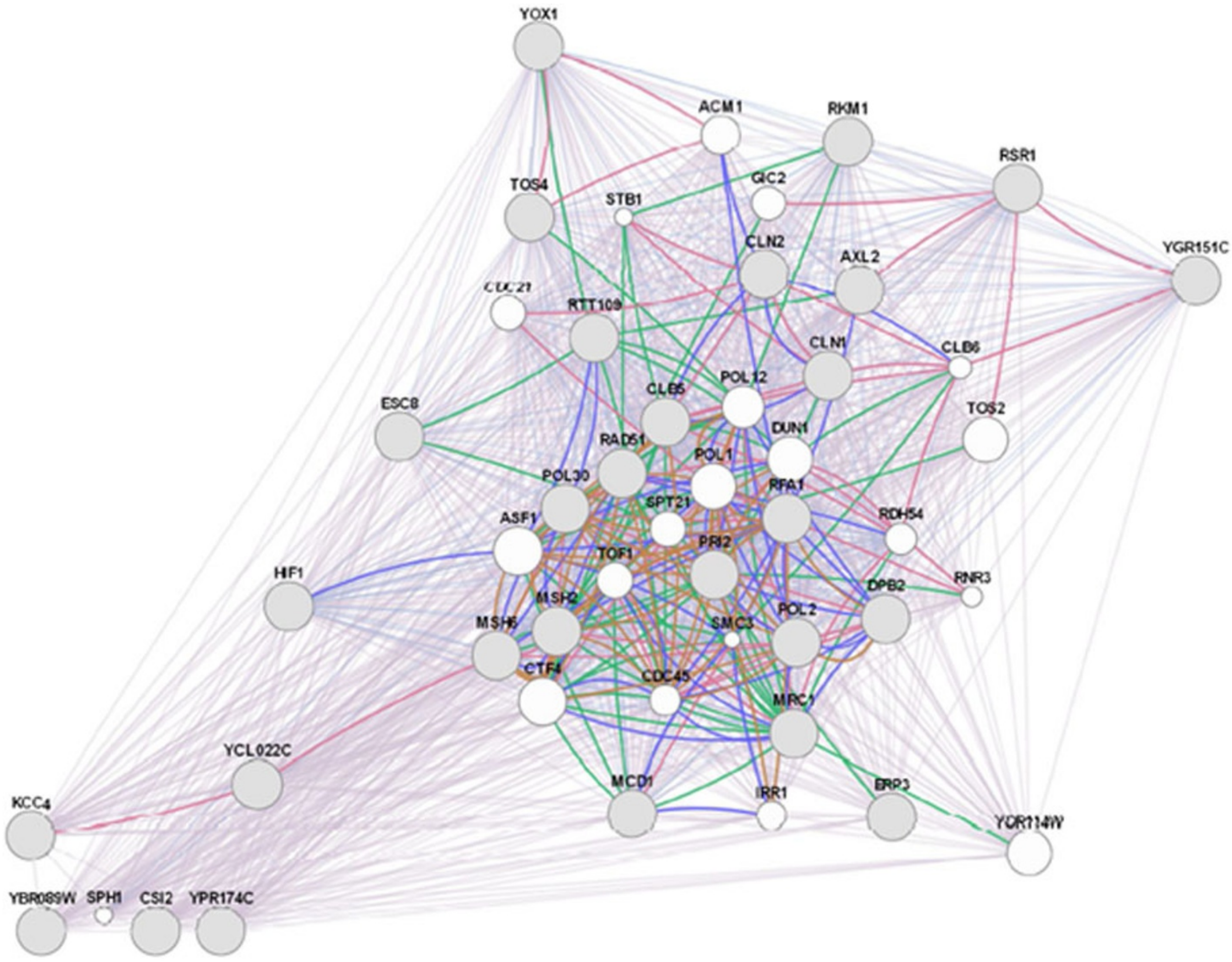
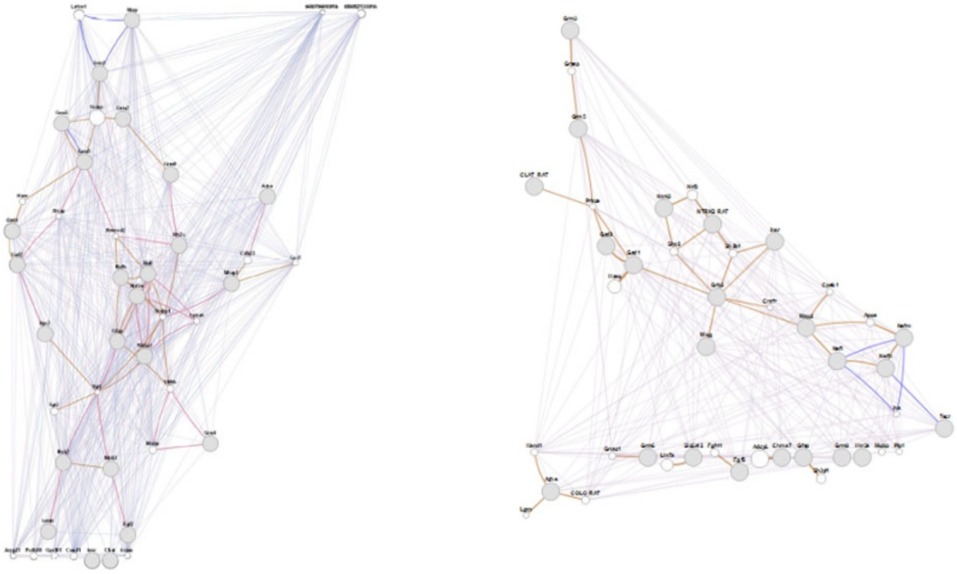


Figure 3: Co-expressed network for Dataset 1

Figure 3 Visualization of co-expressed network The figure3 presents co-expressed network by GeneMANIA for Dataset1.

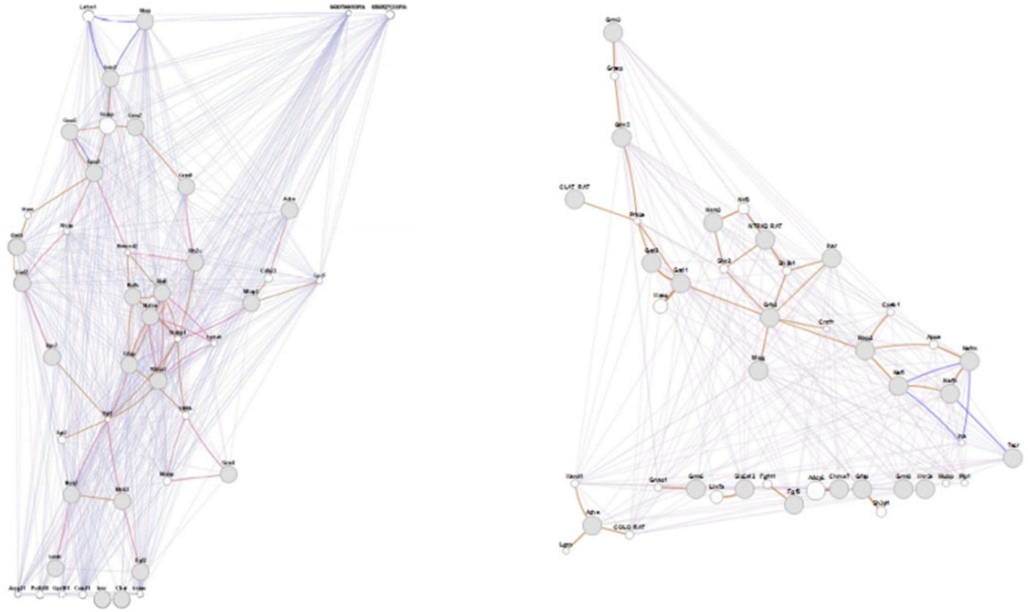


(a) Co expressed network for Dataset 4

(b) Co expressed network for Dataset 5

Figure 4: Co expressed network by GeneMANIA where purple edges represent co-expression, light blue edges represent co localization, green edges represent genetic interactions, dark blue represent physical interactions

Figure 4 Visualization of co-expressed network The figure 4 presents co-expressed network by GeneMANIA for Dataset2 and Dataset3.



(a) Co expressed network for Dataset 4

(b) Co expressed network for Dataset 5

Figure 5: Co expressed network by GeneMANIA where purple edges represent co-expression, light blue edges represent co localization, green edges represent genetic interactions, dark blue represent physical interactions

Figure 5 Visualization of co-expressed network The figure 5 presents co-expressed network by GeneMANIA for Dataset4 and Dataset5.

However, soft thresholding method can be used to construct the adjacency matrix to reduce information loss. Generalized Topological Overlap Measure [25] can be used instead of Topological Overlap Measure to get more accurate results. There is scope to design supervised models to derive gene regulatory network from the co-expression network.

Additional material

Additional file 1: NMRS as a metric This additional file 1 presents the proofs of different metric properties of NMRS measure.

Acknowledgment
This paper is an outcome of a research project supported by (1) DST, Govt. of India in collaboration with ISI, Kolkata and (2) National Science Foundation, USA under grants CNS-095876 and CNS-085173.
This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 13, 2012: Selected articles from The 8th Annual Biotechnology and Bioinformatics Symposium (BIOT-2011). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/13/S13/S1

Author details
[1]Dept. of Comp. Sc. and Engg, Tezpur University, Napaam, Tezpur, India.
[2]Dept. of Computer Science, University of Colorado, Colorado Springs, USA.

Competing interests
The author(s) declare that they have no competing interests.

Published: 24 August 2012

References

1. Wagner A: **How the global structure of protein interaction networks evolves.** *Proc Biol Sci* 2003, **270**:457-466.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:4569-4574.
3. Jeong H, B AL, Mason SP, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41-42.
4. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proceedings. Biological sciences / The Royal Society* 2001, **268(1478)**:1803-1810.
5. Ma H, Zeng AP: **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.** *Bioinformatics* 2003, **19**:270-277.
6. Jeong H, B AL, Mason SP, Oltvai ZN: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.
7. van Noort V, Snel B, Huynen M: **The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model.** *EMBO Reports* 2004, **5(3)**:280-284.
8. Ruan J, Dean A, Zhang W: **A general co-expression network-based approach to gene expression analysis: comparison and applications.** *BMC Systems Biology* 2010, **4**.
9. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:12182-12186 [http://dx.doi.org/10.1073/pnas.220392197].
10. D'Haeseleer P, Liang S, Somogyi R: **Genetic Network Inference: Prom Co-Expression Clustering To Reverse Engineering.** 2000.
11. Butte AJ, Kohane IS, Kohane IS: **Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements.** *Pacific Symposium on Biocomputing* 2000, **5**:415-426.
12. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: Detecting and evaluating dependencies between variables.** *Bioinformatics* 2002, **18**:S231-S240.
13. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P, Alerting E, Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14**:1085-1094.
14. Expression G, Zhu D, Hero AO, Cheng H, Khanna R: **Network constrained clustering for gene microarray data.** *Bioinformatics* 2005, **21**:4014-4021.
15. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**:249-255.
16. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Statistical applications in genetics and molecular biology* 2005, **4**.
17. Yona G, Dirks W, Rahman S, Lin DM: **Effective similarity measures for expression profiles.** *Bioinformatics* 2006, **22(13)**:1616-1622.
18. Jiang D, Tang C, Zhang A: **Cluster Analysis for Gene Expression Data: A Survey.** *IEEE Transactions on Knowledge and Data Engineering* 2004, **16**:1370-1386.
19. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science (New York, N.Y.)* 2002, **297(5586)**:1551-1555.
20. Prim RC: **Shortest connection networks and some generalizations.** *Bell System Technology Journal* 1957, **36**:1389-1401.
21. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nature Genetics* 1999.
22. Berriz GF, King OD, Bryant B, Sander C, Roth FP: **Characterizing gene sets with FuncAssociate.** *Bioinformatics (Oxford, England)* 2003, **19**:2502-2504.
23. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289-300.
24. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Research* 2010, **38**:W214-W220.
25. Yip AM, Horvath S: **Gene network interconnectedness and the generalized topological overlap measure.** *BMC Bioinformatics* 2007, **8**.

doi:10.1186/1471-2105-13-S13-S4
Cite this article as: Mahanta *et al.*: An effective method for network module extraction from microarray data. *BMC Bioinformatics* 2012 13 (Suppl 13):S4.