Netw Model Anal Health Inform Bioinforma
DOI 10.1007/s13721-012-0018-2

ORIGINAL ARTICLE

Module extraction from subspace co-expression networks

Hasin Afzal Ahmed · Priyakshi Mahanta · Dhruba Kr Bhattacharyya · Jugal Kr Kalita

Received: 28 February 2012 / Revised: 30 August 2012 / Accepted: 12 September 2012

Abstract Most existing algorithms for co-expression network construction for the purpose of gene expression data analysis define correlation between a pair of genes over the set of all samples as an edge. In this paper, we propose a way to represent co-expression network that traces correlation among genes over subspace of samples. A method is presented for construction of such a co-expression network. A connectivity measure is also introduced to determine connectivity among genes in the proposed representation of co-expression network. The proposed connectivity measure is used with *k*-means clustering algorithm to extract network modules from the sub-space co-expression network. The methodology has been applied over real life gene expression datasets and the results are validated in terms of external indices such as *p* value and *Q* value.

Keywords Gene expression data analysis · Co-expression network · Multigraph · Connectivity measure · Topological overlap metric · Network module

H. A. Ahmed · P. Mahanta · D. Kr Bhattacharyya (✉)
Department of Computer Science and Engineering,
Tezpur University, Tezpur, India
e-mail: dkb@tezu.ernet.in

H. A. Ahmed
e-mail: hasin@tezu.ernet.in

P. Mahanta
e-mail: priyakshi@tezu.ernet.in

J. Kr Kalita
Department of Computer Science,
University of Colorado, Colorado Springs, CO, USA
e-mail: jkalita@uccs.edu

1 Introduction

Gene expression data analysis enables biologists to narrow the space of probable solutions when performing expensive genomic experiments. This is possible due to the revolutionary microarray technology that monitors expression levels of thousands of genes in a single experiment. A microarray chip is a solid surface of glass, silicon or plastic substrate with probes of DNA or RNA molecules arranged at inter-sectional points in a grid structure. A typical microarray experiment involves preparation of microarray chip, isolation of DNA or RNA molecules from a cell sample, and promotion of hybridization by allowing the extracted DNA or RNA molecules to come in contact with probes. Before hybridization the extracted DNA or RNA solution is dyed with color so that color intensities at the probe spots represent amount of hybridization. Finally, image processing techniques are applied to determine color intensities at different probe spots that represent expression of genes associated with the probes.Analysis of expression data has many applications. Some notable applications are pharmagenomic research and drug discovery, infectious and genetic disease and cancer diagnostics, forensic genetic identification, proteomics and cellular analysis etc (Heller 2002). Data-mining techniques such as classification, clustering, biclustering and triclustering (Jiang et al. 2004; Ahmed et al. 2011a, b; Das et al. 2010; Mahanta et al. 2011) are widely used in gene expression data analysis tasks such as prediction of unknown gene functions, inference of regulatory relationships among genes and disease diagnosis. A number of data-mining techniques have also been proposed specifically for gene expression data analysis. A number of preprocessing tasks such as handling missing values, normalization and feature extraction (Donders et al. 2006; Schadt et al. 2001; Van

Published online: 29 September 2012

Hulse et al. 2012) are performed on gene expression data to make it more effective for different applications.

Gene expression data is usually analyzed to find groups of genes associated with the same biological functions. However, the data produced by microarray experiments hold ample resource for evidence of various biological regulatory relationship among genes as well. Therefore, reverse engineering this expression data to discover these regulatory activities can be very useful for the biologist before conducting in vivo experiments. This reverse engineering task requires representation of the regulatory framework in a computational model. A number of such attempts have been recorded in the literature of gene expression data analysis (De Jong et al. 2002; Zhou et al. 2012). Early researchers tried to model gene regulatory framework considering only pairwise individual interaction among genes by determining pairwise correlations between pairs of genes. Such correlations can be graphically represented in a co-expression network. Gene co-expression networks illustrate associations among genes in terms of their expression similarity and a network-level view of the similarity among a set of genes. Thus, gene co-expression networks provide a good approximation to the complicated web of gene functional associations (Ruan et al. 2010). In co-expression networks, two genes are connected by an undirected edge if their activities have significant association (Lee et al. 2004) computed using gene expression measurements such as Pearson correlation, Spearman correlation and mutual information. A primary objective in designing a co-expression network is to extract highly connected regions from the constructed co-expression network. These regions are often termed network modules. In biological terms, a network module may represent a functional category or a set of co-regulated genes.

In this work, we propose a method to construct co-expression networks reflecting correlations among genes over subspaces of samples and to extract network modules. We also introduce a measure to determine connectivity among genes over subspaces of samples.

1.1 Organization of the paper

The rest of the paper is organized as follows: Section 2 presents related work. Motivation and contributions of the paper are given in Sects. 3 and 4, respectively. The proposed method is described in Sect. 5. Experimental results are provided in Sect. 6. Section 7 presents concluding remarks.

2 Related work

A number of techniques have been proposed for construction of co-expression network in gene expression data analysis. A typical co-expression construction technique accepts a gene expression dataset, computes pairwise correlation among gene expressions and constructs either a weighted co-expression network (where weights are normally the correlation score between the pair of genes) or an unweighted network (where an edge is placed between a pair of nodes if the corresponding gene expressions are correlated with a score more than a threshold). Most of these techniques also provide methods to extract dense network modules which may represent a biologically significant groups of genes. The measures which are frequently used for evaluating correlation in co-expression network analysis are Pearson correlation coefficient, Spearman correlation coefficient and Mutual information. Butte et al. (2000) use Pearson correlation coefficient to place edges among genes in a co-expression network. Spearman correlation coefficient is used as a gene expression similarity measure to construct co-expression networks in D'Haeseleer et al. (1998). Butte and Kohane (2000) and Steuer et al. (2002) reports the use of mutual information to find similarly expressed gene pairs in such networks. While some techniques operate on the adjacency matrices of networks in order to partition network nodes into groups (Lee et al. 2004; Zhu et al. 2005), other techniques rely on special purpose algorithms for identifying subnetworks with certain properties (Stuart et al. 2003). The method called Qcut (Ruan et al. 2010) constructs co-expression networks using rank-based Pearson correlation between a node and rest of the nodes. It also optimizes modularity, which is an objective function defined as the difference between percentage of intra-community edges and random expectation. Table 1 lists different co-expression network construction and module extraction techniques.

3 Motivation

Our literature review finds that the existing techniques of gene co-expression network construction consider all samples when determining correlations among genes. But, it is well known in molecular biology that only a small subset of the genes participates in any cellular process and that any cellular process takes place only in a subset of the samples (Jiang et al. 2004). The fact can be easily understood if we look at the origin of the gene expression data. Gene expression data is the outcome of microarray experiments that measure expression levels of genes in a number of samples. In applications such as function prediction of genes and disease diagnosis, we are not only interested in evaluating correlation between a pair of genes over whole set of samples because a biological process may occur only in a subset of samples. To elucidate this statement let us consider the two expression patterns

Table 1 A comparison of some existing co-expression network construction techniques

Method	Approach of module extraction	Type of network	Measure used
Butte et al. (2000)	Connected nodes are extracted as relevance network	Unweighted	Pearson's correlation
Butte and Kohane (2000)	Connected nodes are extracted as relevance network	Unweighted	Mutual information
Lee et al. (2004)	Hierarchical clustering is applied on interaction matrix	Weighted	Pearson's correlation
Zhu et al. (2005)	Hierarchical clustering is applied on shortest path matrix derived from Giant Connected Component	Weighted	FDR and MAS
Stuart et al. (2003)	*K* means clustering is applied on weighted matrix of *p* values of correlation across multiple species	Weighted	Pearson's correlation
Ruan et al. (2010)	Extracts components of graph with more intra-component edges	Unweighted	Pearson's correlation

FDR False Discovery Rate, *MAS* Minimum Acceptable Strength

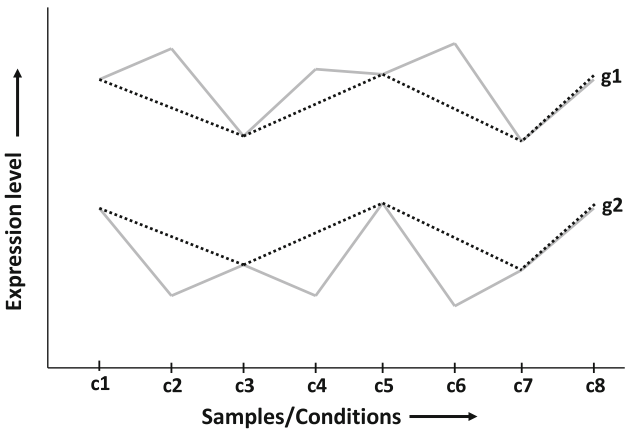


Fig. 1 Two gene expression patterns with correlation over subset of samples {*c1*, *c3*, *c5*, *c7*, *c8*}

corresponding to two genes g1 and g2 as shown in Fig. 1. The figure plots expression patterns of g1 and g2 (solid lines) over set of samples c1, c2, c3, c4, c5, c6, c7 and c8. Now let us consider expression patterns of the genes over subset of samples c1, c3, c5, c7 and c8 (dotted line). Correlations over subspace of samples are normally ignored by most co-expression network construction techniques, and evaluate correlation over whole sets of samples. But as per requirements of certain applications, these patterns are of prime importance. Therefore, there is a need to incorporate subspace correlations rather than full-space correlations when constructing co-expression network for more biologically relevant results.

Most correlation measures are sensitive to the order of samples in the gene expression matrix. But if we analyze the gene sample microarray data, we see that there is no explicit ordering of samples. So a correlation measure should be robust to the problem of order sensitivity. This problem can be avoided by operating on each possible pair of samples instead of operating only on consecutive pairs of samples in the gene expression data. Another problem in the process of determining the correlation between two expression patterns is the presence of noisy data. A noisy expression value of a gene against a sample should not affect the correlation across the rest of the samples.

4 Contributions

Our work is aimed to design a framework for construction of co-expression networks that takes into account correlations among genes over subsets of samples. The work also proposes a connectivity measure that can be used with subspaces among genes when constructing gene co-expression networks. This paper makes the following contributions to gene expression data analysis.

1. We propose a representation for co-expression networks that can handle correlations among genes over subsets of samples. We use the term *subspace co-expression network* to refer to such a network. A subspace co-expression network can be graphically viewed as a multigraph. Multiple edges between two nodes representing a pair of genes in the multigraph correspond to different subsets of samples across which the genes are correlated.
2. We propose a TOM (Ravasz et al. 2002)-based connectivity measure named TSOM (Topological Subspace Overlap Metric) which can be used to determine connectivity between a pair of genes in the proposed subspace co-expression network.
3. We also introduce an unsupervised module extraction technique from the subspace co-expression network based on (1) TSOM measure and (2) *k*-means clustering algorithm.

5 Method

When we consider a set of samples, a pair of genes can be correlated in more than one subset of samples. Considering

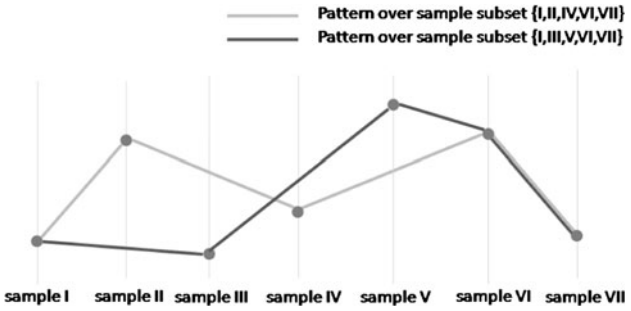


Fig. 2 Two ordered sample subsets of a gene expression pattern

this fact, to represent a gene co-expression network, i.e. a graph $G(V, E)$, where V is the set of vertices or genes and E is the set of edges, and between a pair of vertices $\{V_i, V_j\}$, there can be multiple edges corresponding to multiple subspaces of samples. we use multigraph to represent gene co-expression network where multiple edges between a pair of genes or nodes correspond to different subspaces of samples. We consider two genes to be correlated for a pair of samples if *arctan* of the angles formed by the expression values for the sample pair do not differ by more than a threshold. This threshold is computed as a function of standard deviation of *arctan* of expression values for all the genes over the pair of samples. Per our notion of correlation, a pair of genes can be correlated over different subsets of samples. For example the gene expression pattern in Fig. 2 can be correlated with another gene separately in two subsets of samples I, II, IV, VI, VII and I, III, V, VI, VII. Figure 2 plots patterns formed by expression values of the gene over these two subsets of samples from the set of possible subsets of samples. This point onwards, we will use the term *ordered sample subset* to refer to such a subset of samples.

The symbols provided in Table 2 and the definitions given next are used to describe the proposed method.

Definition 1 A *sample subspace* for a gene expression dataset is defined as a subset of samples or conditions associated with the gene expression dataset. We use the term subspace to refer the possible sample subspaces in a gene expression dataset.

Definition 2 A *subspace co-expression network* is a co-expression network with the ability to convey information about the subset of samples over which genes or nodes are correlated. A subspace co-expression network can be represented by a multigraph $G = (V, E)$ where V is a set of vertices that represent genes and E is a set consisting of unordered pairs of vertices each of which is associated with a subset of samples.

Table 2 Symbolic representations

Symbol	Meaning
G	Gene expression matrix
$G(i,j)$	Expression value of ith gene for jth sample
g_i	Expression value of ith gene in G
λ	Deviation ratio
D	Discretized gene expression matrix
$D_i^{m,n}$	Discretized value assigned to g_i for sample pair {m, n}
C_{ij}^k	Set of samples corresponding to kth edge over which ith and jth genes are correlated
$\beta(p, q)$	Standard deviation of the distribution $D_1^{p,q}, D_2^{p,q}, \ldots, D_m^{p,q}$

Definition 3 An *adaptively discretized matrix* D is a discretized form of expression matrix G of order $m \times (n \times (n-1))$, where m is the number of genes in G and n is the number of samples in G such that $D_i^{p,q}$ and $D_j^{p,q}$ are assigned the same discrete value if $|arctan(G(i, p) - G(i, q)) - arctan(G(j, p) - G(j, q))| \leq \lambda \times \beta(p, q)$, where λ is a user defined threshold named deviation ratio.

Definition 4 Two genes g_i and g_j are said to be correlated over an *ordered sample subset* $\{s_1, s_2, \ldots, s_n\}$ if $D_i^{s_k, s_{k-1}} = D_j^{s_k, s_{k-1}}$, $k = 2, 3, \ldots, n$

A block diagram of our method is presented in Fig. 3. The expression matrix is discretized considering angles formed by expression values of genes for each pair of samples. These discretized expression data are used to form a multigraph that represents the subspace co-expression network. Then pairwise TSOM values among genes are computed from the multigraph. These connectivity scores are stored in the form of a matrix. Finally, K-means (Hartigan and Wong 1979) algorithm is employed to extract network modules using the connectivity matrix.

5.1 Adaptive discretization

In this work, we adaptively discretize the expression matrix of order $m \times n$ to a discrete matrix of order $m \times \frac{n \times (n-1)}{2}$. In the discretization process, for each pair of conditions, the angle formed by the expression values corresponding to a pair of conditions is computed as arctan of the difference of the expression values, as presented in Fig. 4. For two genes g_i and g_j, arctan value for pth and qth conditions is computed as,

$$\theta = arctan(G(i,p) - G(i,q))$$

These arctan values are then discretized based on a threshold by assigning all the values close by an amount

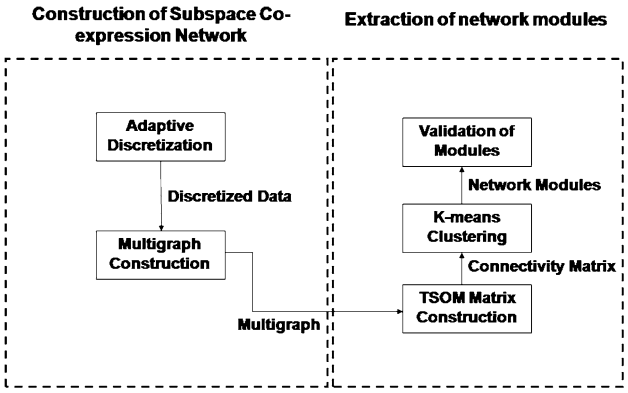


Fig. 3 Proposed methods of co-expression network construction and module extraction

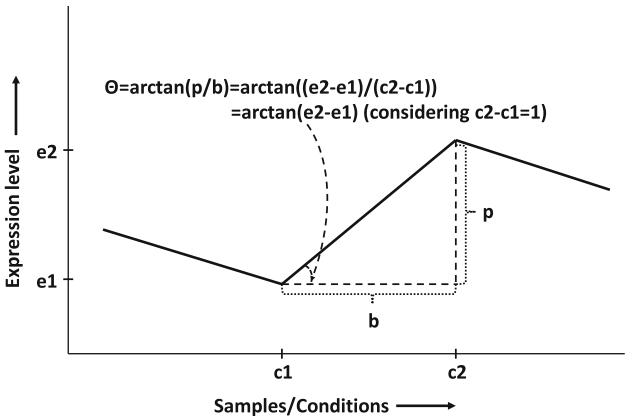


Fig. 4 Computing angle formed by a pair of expression values of a gene corresponding to a pair of conditions

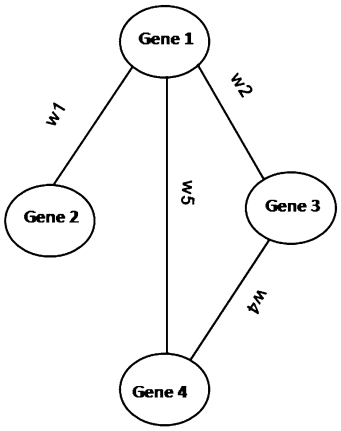
less than this threshold the same discrete value. This threshold is computed as a function of standard deviation of the values. The discretization process is presented next in details.

Algorithm Adap-discretize()
INPUT: Expression matrix, G, Deviation Ratio, λ
OUTPUT: Discretized matrix, D
Steps:
For each pair of conditions c_i and c_j in G

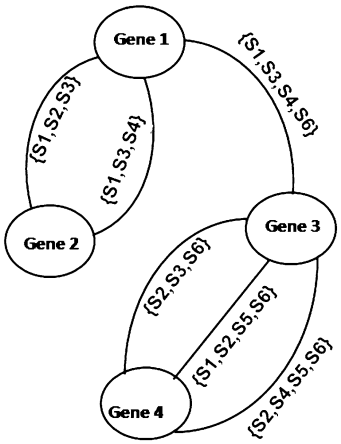
- Convert each pair of expression values corresponding to two conditions of genes to its *arctan* form.
- Sort these *arctan* values.
- Group the sorted values by placing the values that are close to each other with a distance $<\lambda \times \beta$ in the same group, where β is the standard deviation.
- Assign all the members of each group a unique alphabet.

5.2 Subspace co-expression network construction

The subspace co-expression network is an enhancement of the traditional co-expression network. This network can



(a) Traditional co-expression network



(b) Subspace co-expression network

Fig. 5 Traditional versus subspace co-expression network

trace the subsets of samples along which a pair of genes are correlated. Nodes in this network represent genes in the expression data. We use the terms *nodes* and *genes* interchangeably in this paper. Each edge between a pair of nodes (genes) is labeled with a subset of samples. Per angular similarity that is used to define correlation between two genes, there can be multiple subsets under which a pair of genes are correlated. Unlike traditional co-expression networks, where there can be only one edge between a pair of genes or nodes, subspace co-expression network may contain more than one edge between a pair of genes or nodes as shown in Fig. 5a, b. An edge represents a subset of samples over which the vertices are correlated. This multigraph is constructed as follows:

- For each pair of genes g_i and g_j,
- Compare corresponding entries in D to find the pairs of samples under which g_i and g_j have the same discrete value.
- Process the pairs to derive the ordered sample subsets. For example, if (1, 5), (3, 5) and (5, 6) are the similar

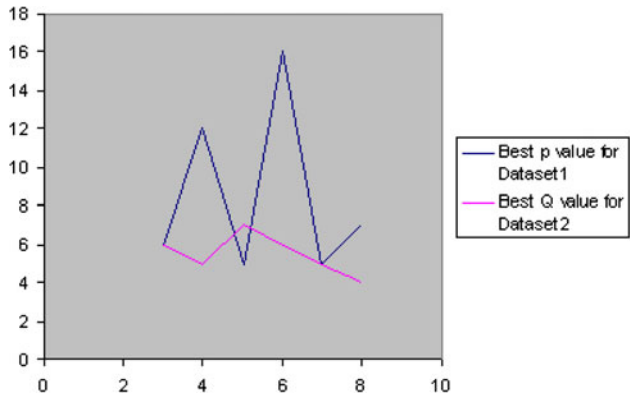


Fig. 6 Determination of k value for extraction of network modules

sample pairs, the derived ordered sample subsets are 1, 5, 6 and 3, 5 ,6.

– For each ordered sample subset, draw an edge between the nodes corresponding to g_i and g_j with the sample subset as its label.

5.3 Topological subspace overlap metric

The topological overlap metric (Ravasz et al. 2002) is a similarity measure, which is useful in biological networks. This measure is generally defined for weighted and unweighted networks where there is a single edge between a pair of nodes. For unweighted networks (i.e., $a_{ij} = 1$ or $= 0$), the topological overlap matrix is defined by (Ravasz et al. 2002),

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

where $l_{ij} = \sum_u a_{iu} a_{ij}$ is the number of common neighbors and $k_i = \sum_u a_{iu}$ is the node connectivity.

TOM can be used to determine connectivity in a network where there can be at most one edge between a pair of nodes. So this measure cannot be applied on a subspace co-expression network which may have multiple edges between a pair of nodes corresponding to different subsets of samples over which genes or nodes are correlated. We propose a TOM-based connectivity measure named TSOM (Topological Subspace Overlap Metric) which can be applied on a subspace co-expression network .

The Topological Subspace Overlap Metric between ith and jth nodes in a network with n nodes is defined as,

$$TSOM_{ij} = 0.5 \times \eta + 0.5 \times \omega \quad (2)$$

where

Neighbourhood similarity, $\eta = \frac{\sum_{u=1}^{n} |C_{iu}^{p} \cap C_{uj}^{q}|}{\sum_{u=1}^{n} min(|C_{iu}^{p}|, |C_{uj}^{q}|)}$, $1 \leq p \leq n_{iu}$, $1 \leq q \leq n_{uj}$ such that $|C_{iu}^{p} \cap C_{uj}^{q}|$ is maximum and $|C_{iu}^{p} \cap C_{uj}^{q}| > 2$,

Direct connectivity, $\omega = \frac{|C_{ij}^{r}|}{s}$, $1 \leq r \leq n_{ij}$ such that $|C_{ij}^{r}|$ is maximum,

Table 3 Datasets used for evaluation

Serial no.	Dataset	No. of genes/ No. of conditions	Source
1	Rat CNS	112/9	http://cmgm.standford.edu/pbrown/sporulation
2	*Arabidopsis thaliana*	138/8	http://homes.esat.kuleuven.be/thijs/Work/Clustering.html
3	Subset of yeast cell cycle	384/17	http://faculty.washington.edu/kayee/cluster

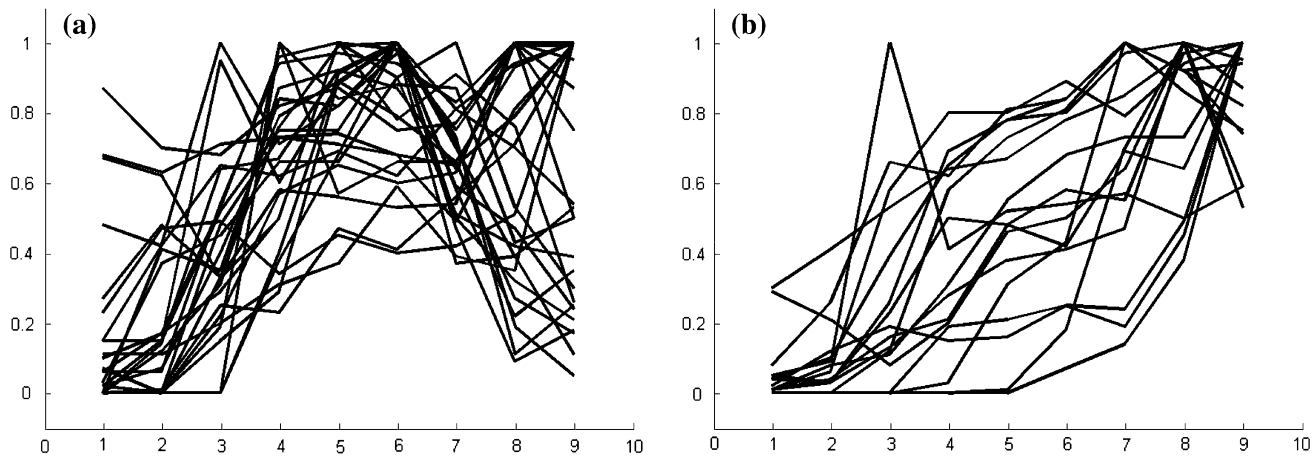


Fig. 7 Visualization of different network modules of Dataset 1

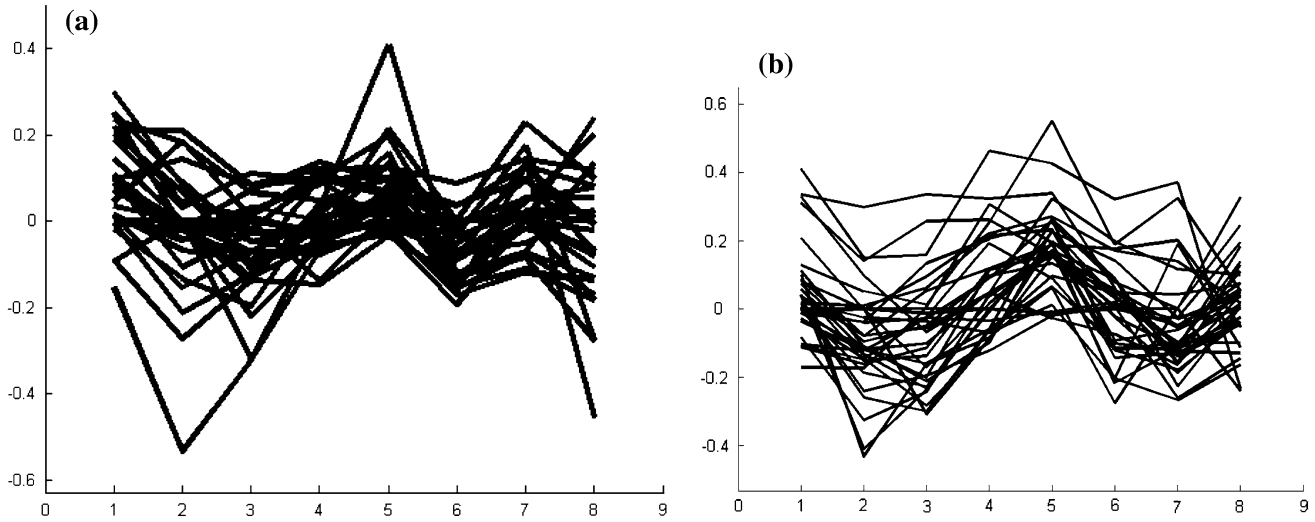


Fig. 8 Visualization of different network modules of Dataset 2

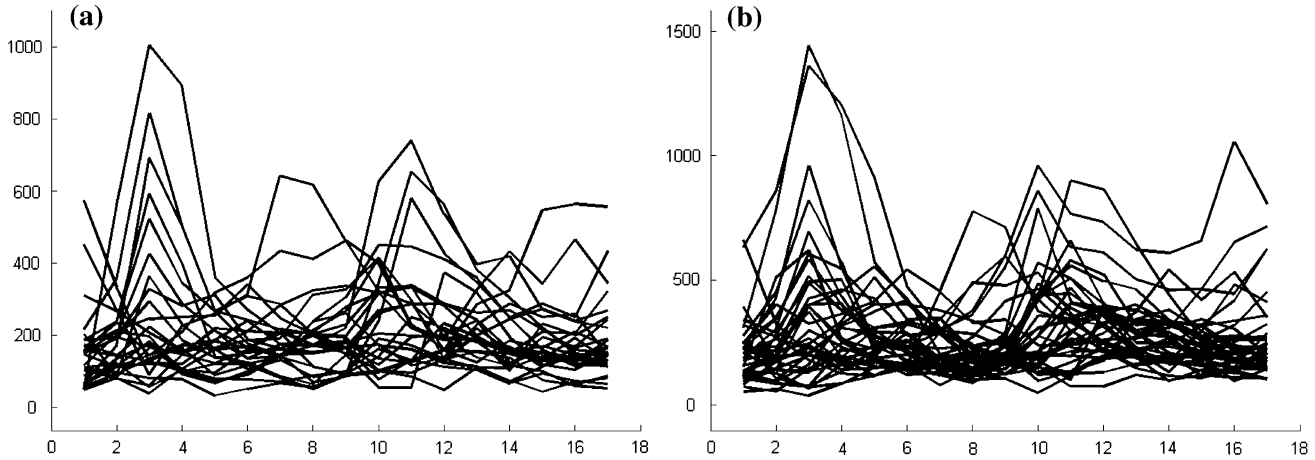


Fig. 9 Visualization of different network modules of Dataset 3

Table 4 *p* value of one of the network modules of Dataset 1

p value	GO number	GO category
9.64E−06	GO:0060267	Positive regulation of respiratory burst
4.35E−06	GO:0008291	Acetylcholine metabolic process
1.45E−05	GO:0033265	Choline binding
4.82E−07	GO:0008083	Growth factor activity

n_{ij} is the total number of edges between ith and jth nodes, and

s is the number of samples in G.

When computing connectivity between a pair of nodes, we consider two components viz., neighbourhood similarity η and direct connectivity ω. Direct connectivity takes into account the edge between nodes corresponding to the genes, g_i and g_j while neighbourhood similarity takes into account the the nodes which are connected to both the nodes corresponding to genes g_i and g_j. To avoid biasing, we have assigned equal weights (i.e., 0.5) to both the components η and ω.

5.4 Extraction of network modules

To extract modules from the subspace co-expression network, pairwise TSOM scores of the genes or nodes are computed. Using these connectivity scores as similarity values, the k-means (Hartigan and Wong 1979) clustering algorithm is applied to produce clusters. The k-means

Table 5 *p* value of some of the network modules of Dataset 3

Network modules	*p* value	GO number	GO category
Module 1	5.09E−06	GO:0007346	Regulation of mitotic cell cycle
	1.22E−07	GO:0010564	Regulation of cell cycle process
	9.72E−10	GO:0007049	Cell cycle
	3.59E−05	GO:0007067	Mitosis
	2.17E−06	GO:0051726	Regulation of cell cycle
	3.91E−05	GO:0000280	Nuclear division
	1.01E−06	GO:0051301	Cell division
	5.65E−05	GO:0048285	Organelle fission
	4.68E−09	GO:0022402	Cell cycle process
	1.09E−05	GO:0044454	Nuclear chromosome part
	2.38E−05	GO:0022403	Cell cycle phase
	1.28E−05	GO:0044427	Chromosomal part
	5.35E−05	GO:0044428	Nuclear part
Module 2	4.27E−08	GO:0006302	Double-strand break repair
	1.85E−07	GO:0005935	Cellular bud neck
	1.69E−05	GO:0005576	Extracellular region
	2.525E−05	GO:0006260	DNA replication
	9.611E−08	GO:0051726	Regulation of cell cycle
	6.12E−09	GO:0051301	Cell division
	8.10E−11	GO:0007049	Cell cycle
	6.15E−06	GO:0010564	Regulation of cell cycle process
	4.61E−11	GO:0022402	Cell cycle process
	1.87E−06	GO:0030427	Site of polarized growth
	1.10E−05	GO:0044454	Nuclear chromosome part

clustering algorithm iteratively partitions set of objects into groups of similar objects. The algorithm starts with n initial seeds and assigns the rest of the objects to one of these seeds. After the assignment, centroids of the partial clusters are computed as seeds for the next iteration. The process is repeated until the centroids do not change. The clusters extracted by k-means algorithm are considered extracted network modules. Discovery of these modules involve the following steps.

- Compute pairwise TSOM scores of the genes to generate a connectivity matrix *Con* of order $m \times n$, where m is the number of genes and n is the number of samples in G.
- Subtract each similarity value in *Con* from 1 to construct distance matrix *Dist*.
- Feed *Dist* is then fed to k-means clustering algorithm to obtain the clusters. These clusters actually represent relatively dense regions in the network and are extracted network modules. To determine the value of k, we tried different possible values of k and choose the one with highest biological significance(i.e. lowest p or Q value) as shown in Fig. 6.

6 Experimental results

We implement the method in MATLAB and test it on three benchmark microarray datasets given in Table 3. The test platform is a Sun workstation with Intel(R) Xenon(R) 3.33 GHz processor and 6 GB memory running Windows XP operating system.

6.1 Validation

The performance of the algorithm on the publicly available benchmark microarray dataset is measured in terms of p value (Tavazoie et al. 1999) and Q value (Benjamini and Hochberg 1995). Some of the network modules extracted by our method are visually presented in Fig. 7a, b for Dataset 1 and Fig. 8a, b for Dataset 2 and Fig. 9a, b for Dataset 3.

6.1.1 p value

We evaluate biological significance of the extracted network modules using p value (Tavazoie et al. 1999). p value

Table 6 Q value of some of the network modules of Dataset 1

Network modules	Go annotation	Q value
Module 1	Actor binding	4.658757E−10
	Negative regulation of neuron apoptosis	4.05127E−6
	Growth factor receptor binding	4.05127E−6
	Regulation of platelet activation	4.369925E−6
	Cytokine receptor binding	2.655455E−5
	Negative regulation of blood coagulation	3.581839E−5
	Negative regulation of coagulation	3.738841E−5
	Peptidyl-tyrosine modification	3.738841E−5
	Regulation of neuron apoptosis	3.738841E−5
	Transmembrane receptor protein tyrosine kinase signaling pathway	3.738841E−5
	Peptidyl-tyrosine phosphorylation	3.738841E−5
	Neuron apoptosis	5.419012E−5
	Neuron death	6.108174E−5
	Cellular response to growth factor stimulus	6.251706E−5
Module 2	O-acyltransferase activity	3.959642E−6
	Intermediate filament organization	2.29248E−5
	Regulation of neuron projection development	3.349611E−5
	Regulation of cell projection organization	7.442948E−5
	Intermediate filament cytoskeleton organization	9.65029E−5
Module 3	Small molecule catabolic process	1.236104E−7
	Response to cadmium ion	1.236104E−7
	Apoplast	4.255192E−6
	Oxidation–reduction process	2.666989E−5
	Extracellular region	2.666989E−5
	Copper ion binding	2.682103E−5
	Monosaccharide catabolic process	2.703877E−5
	Glucose catabolic process	2.703877E−5
	Hexose catabolic process	2.703877E−5
	Alcohol catabolic process	3.247549E−5
	Glucose metabolic process	3.877089E−5

signifies how well the genes in a network module match various GO categories. A low p value for the set of genes indicates that the genes belong to enriched functional categories that are biologically significant. A cumulative hypergeometric distribution is used to compute the p value. For a given GO category, the probability p of getting k or more genes within a cluster of size n, is defined as (Berriz et al. 2003),

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} \tag{3}$$

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

To compute p value, we use the Web-based tool called FuncAssociate (Berriz et al. 2003). FuncAssociate uses Molecular Function and Biological Process annotations in Gene Ontology to compute the hyper-geometric functional enrichment score. The enriched functional categories for some of the network modules obtained by our algorithm on the datasets are presented in Tables 4 and 5. The network modules produced by the method for these datasets contain the highly enriched functional categories of growth factor activity, regulation of cell cycle, cell cycle process, cell division, cellular bud neck, cell division, cell cycle with p values of $4.82\times\ 10^{-07}$, 2.17×10^{-06}, 4.68×10^{-09}, 1.01×10^{-06}, 1.85×10^{-07}, 6.12×10^{-09}, $8.10\ \times 10^{-11}$ respectively, being the highly enriched GO categories. Table 9 presents some of the functional categories which are better or equally detected by our technique as compared to Qcut (Ruan et al. 2010) technique for Dataset 3 in terms of p value.

6.1.2 Q value

The Q value (Benjamini and Hochberg 1995) for a set of genes is the proportion of false positives among all genes

Table 7 Q value of some of the network modules of Dataset 2

Network modules	GO annotation	Q value
Module 1	Carboxy-lyase activity	3.663364E−7
	Carbon–carbon lyase activity	1.401449E−6
	Positive regulation of cellular carbohydrate metabolic process	4.156939E−6
	Positive regulation of carbohydrate metabolic process	4.156939E−6
	Insulin-like growth factor receptor binding	5.726434E−6
	Positive regulation of glycogen biosynthetic process	5.726434E−6
	Positive regulation of glycogen metabolic process	8.678437E−6
	Lyase activity	8.678437E−6
	Regulation of carbohydrate biosynthetic process	1.106793E−5
	Insulin receptor binding	2.672179E−5
	Regulation of glucan biosynthetic process	4.677805E−5
	Regulation of polysaccharide biosynthetic process	4.677805E−5
	Regulation of glycogen biosynthetic process	4.677805E−5
	Regulation of polysaccharide metabolic process	5.419073E−5
	Positive regulation of glucose metabolic process	5.419073E−5
	Regulation of glycogen metabolic process	5.419073E−5
	Regulation of cellular carbohydrate metabolic process	8.038949E−5
	Regulation of carbohydrate metabolic process	8.674458E−5
	Positive regulation of nuclear division	9.346483E−5
	Positive regulation of mitosis	9.346483E−5
	Glucan biosynthetic process	9.746132E−5
	Glycogen biosynthetic process	9.746132E−5
Module 2	Secondary metabolic process	9.03893E−16
	Response to wounding	2.618558E−11
	Toxin catabolic process	2.894879E−10
	Toxin metabolic process	2.894879E−10
	Glutathione transferase activity	3.313905E−10
	Response to bacterium	1.901872E−9
	Transferase activity, transferring alkyl or aryl (other than methyl) groups	2.015514E−9
	Defense response to bacterium	2.866611E−7
	Aromatic compound biosynthetic process	5.079959E−7
	Glutathione binding	9.067702E−7
	Modified amino acid binding	9.067702E−7
	Peptide binding	9.363736E−6
	Tryptophan metabolic process	2.01138E−5

that are as or more extremely differentially expressed. GeneMANIA (Warde-Farley et al. 2010) reports GO categories and Q values from an FDR (False Discovery Rate) corrected hypergeometric test. Q values are estimated using the Benjamini Hochberg procedure (1995). Different GO categories of the co-expression networks produced by the method are displayed up to a Q value cutoff of 0.1 in Tables 6, 7 and 8. The co-expression network modules produced by the method contain the highly enriched functional catagories of actor binding, negative regulation of neuron apoptosis, regulation of platelet activation, O-acyltransferase activity, carboxy-lyase activity, secondary metabolic process, response to wounding, toxin catabolic process with Q values of 4.6×10^{-10}, 4.05×10^{-6}, 4.36×10^{-6}, 3.95×10^{-6}, 3.6×10^{-7}, 9.03×10^{-16}, 2.61×10^{-11}, 2.89×10^{-10}, respectively, being the highly enriched GO categories. From the results of p and Q values, we can conclude that our method shows a good enrichment of functional categories and therefore is able to discover modules with a good biological significance. Table 10 presents some of the functional categories which are better or equally detected by our technique as compared to Qcut (Ruan et al. 2010) technique for Dataset 3 in terms of Q value.

Table 8 Q value of some of the network modules of Dataset 3

Network modules	GO annotation	Q value
Module 1	Cellular bud	1.745636E−22
	Cellular bud neck	1.052997E−21
	Site of polarized growth	6.084308E−21
	Cell cortex	2.165208E−11
	Cell division	4.041012E−11
	Cell division site	1.437293E−10
	Cell division site part	1.437293E−10
	Cytokinesis	1.465696E−9
	Cytoskeletal part	2.03721E−9
	Cell cortex part	2.08987E−9
	Cytoskeleton	2.136264E−9
	Cellular bud neck contractile ring	6.60869E−9
	Contractile ring	9.880097E−9
	Actomyosin contractile ring	9.880097E−9
Module 2	Interphase	7.875588E−10
	Interphase of mitotic cell cycle	7.875588E−10
	Mitotic cell cycle	7.875588E−10
	S phase of mitotic cell cycle	7.875588E−10
	Nuclear replication fork	1.471291E−9
	S phase	3.742464E−9
	Replication fork	9.535708E−9
	DNA-dependent DNA replication	2.229168E−8
	DNA replication	3.191653E−8
	DNA-dependent DNA replication initiation	3.793542E−8
	Replication fork protection complex	4.194858E−7
	Pre-replicative complex	5.367278E−7
	Protein-DNA complex	5.367278E−7
	Pre-replicative complex assembly	7.924449E−7
	DNA recombination	9.360559E−7
	Cellular bud neck	2.977326E−6
	DNA strand elongation involved in DNA replication	3.566514E−6

Table 9 Comparison of results of Qcut and proposed technique for Dataset 3 in terms of p value

Annotation	GO Term	Proposed technique	Qcut
Regulation of cytokinesis	GO:0032465	5.86E−05	3.4E−05
Inner plaque of spindle pole body	GO:0005822	5.21E−07	7.79E−06
Microtubule organizing center part	GO:0044450	4.58E−05	1.87E−05
DNA geometric change	GO:0032392	1.99E−05	2.52E−05
Regulation of protein serine/threonine kinase activity	GO:0071900	4.61E−05	1.45E−05
Regulation of cell cycle	GO:0051726	9.61E−08	1.54E−07
Cell division	GO:0051301	6.13E−09	2.77E−09
Regulation of cell cycle process	GO:0010564	2.35E−07	4.45E−07
Nuclear part	GO:0044428	5.35E−05	3.95E−05

Table 10 Comparison of results of Qcut and proposed technique for Dataset 3 in terms of Q value

Annotation	GO Term	Proposed technique	Qcut
Cellular bud	GO:0005933	1.74E−22	2.51E−07
Cellular bud neck	GO:0005935	1.05E−21	1.23E−05
Site of polarized growth	GO:0030427	6.08E−21	2.10E−05
Cell division	GO:0051301	4.04E−11	1.54E−03
Cell division site	GO:0032153	1.44E−10	5.77E−03
Cytokinesis	GO:0000910	1.46E−09	2.10E−04
Cellular bud neck contractile ring	GO:0000142	6.61E−09	2.70E−04
Contractile ring	GO:0070938	9.88E−09	3.60E−04
Actomyosin contractile ring	GO:0005826	9.88E−9	3.60E−04
Interphase	GO:0051325	7.87E−10	2.92E−07
Interphase of mitotic cell cycle	GO:0051329	7.87E−10	2.42E−07
S phase of mitotic cell cycle	GO:0000084	7.87E−10	9.91E−05
Replication fork protection complex	GO:0031298	4.19E−07	1.39E−04
Pre-replicative complex assembly	GO:0006267	7.92E−7	4.46E−05

7 Conclusions

In this paper, we have proposed a method to construct co-expression network over subspaces of samples. A connectivity measure has also been proposed to evaluate connectivity between a pair of nodes in the proposed network structure. We use k-means clustering to extract network modules from the generated subspace co-expression network. We validate extracted modules from multiple real life datasets using p value and Q value. The results are highly satisfactory.

Acknowledgments This work is supported by DST, Government of India through INSPIRE program. The work is also an outcome of a research project in collaboration with CSCR, ISI, Kolkata funded by DST, Government of India.

References

Ahmed H, Mahanta P, Bhattacharyya D, Kalita J (2011a) Gerc: tree based clustering for gene expression data. In: 2011 IEEE 11th international conference on Bioinformatics and Bioengineering (BIBE), pp 299–302. IEEE, New York.

Ahmed H, Mahanta P, Bhattacharyya D, Kalita J, Ghosh A (2011b) Intersected coexpressed subcube miner: An effective triclustering algorithm. In: 2011 World Congress on Information and Communication Technologies (WICT), pp 846–851. IEEE, New York

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300

Berriz G, King O, Bryant B, Sander C, Roth F (2003) Characterizing gene sets with funcassociate. Bioinformatics 19(18):2502–2504

Butte A, Kohane I (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 5:418–429

Butte A, Tamayo P, Slonim D, Golub T, Kohane I (2000) Discoveringfunctional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proc Nat Acad Sci 97(22):12182–12186

Das R, Bhattacharyya D, Kalita J (2010) Clustering gene expression data using an effective dissimilarity measure. Int J Comput BioSci 1(1):55–68

De Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 9(1):67–103

D'Haeseleer P, Wen X, Fuhrman S, Somogyi R (1998) Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In: Second international workshop on information processing in cells and tissues, pp 203–212

Donders A, van der Heijden G, Stijnen T, Moons K (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59(10):1087–1091

Hartigan J, Wong M (1979) Algorithm as 136: A k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):100–108

Heller M (2002) Dna microarray technology: devices, systems, and applications. Ann Rev Biomed Eng 4(1):129–153

Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng 16(11):1370–1386

Lee H, Hsu A, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14(6):1085–1094

Mahanta P, Ahmed H, Bhattacharyya D, Kalita J (2011) Triclustering in gene expression data analysis: a selected survey. In: 2011 2nd National Conference on Emerging trends and applications in computer science (NCETACS), pp 1–6. IEEE, New York

Ravasz E, Somera A, Mongru D, Oltvai Z, Barabási A (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Ruan J, Dean A, Zhang W (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol 4(1):8

Schadt E, Li C, Ellis B, Wong W (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem 84(S37):120–125

Steuer R, Kurths J, Daub C, Weise J, Selbig J (2002) The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18(suppl 2):S231–S240

Stuart J, Segal E, Koller D, Kim S (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255

Tavazoie S, Hughes J, Campbell M, Cho R, Church G et al (1999) Systematic determination of genetic network architecture. Nat Genet 22:281–285

Van Hulse J, Khoshgoftaar T, Napolitano A, Wald R (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. Netw Model Anal Health Inform Bioinforma 1(1–2):1–15

Warde-Farley D, Donaldson S, Comes O, Zuberi K, BadrawiR, Chao P, Franz M, Grouios C, Kazi F, Lopes C et al (2010) The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res Suppl 38(suppl 2):W214–W220

Zhou Y, Qureshi R, Sacan A (2012) Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression. Netw Model Anal Health Inform Bioinforma 1(1–2):1–15

Zhu D, Hero A, Cheng H, Khanna R, Swaroop A (2005) Network constrained clustering for gene microarray data. Bioinformatics 21(21):4014–4020