FUMET: A fuzzy network module extraction technique for gene expression data

PRIYAKSHI MAHANTA[1], HASIN AFZAL AHMED[1], DHRUBA KUMAR BHATTACHARYYA[1] and ASHISH GHOSH[2]

[1]*Department of Computer Science and Engineering, Tezpur University, Napaam 784 028, India*

[2]*Machine Intelligent Unit, Indian Statistical Institute, Kolkata 700 108, India*

(Emails, PM – priyakshi@tezu.ernet.in, HAA – hasin@tezu.ernet.in, DKB – dkb@tezu.ernet.in, AG – ash@isical.ac.in)

Construction of co-expression network and extraction of network modules have been an appealing area of bioinformatics research. This article presents a co-expression network construction and a biologically relevant network module extraction technique based on fuzzy set theoretic approach. The technique is able to handle both positive and negative correlations among genes. The constructed network for some benchmark gene expression datasets have been validated using topological internal and external measures. The effectiveness of network module extraction technique has been established in terms of well-known p-value, Q-value and topological statistics.

[Mahanta P, Ahmed HA, Bhattacharyya DK and Ghosh A 2014 FUMET: A fuzzy network module extractiontechnique for gene expression data. *J. Biosci.* **39** 351–364] **DOI** 10.1007/s12038-014-9423-2

1. Introduction

Large amount of gene expression data produced by revolutionary microarray technology have given rise to a number of challenges for the researchers in the field of gene expression data analysis. These gene expression datasets are evidences of a number of biological phenomena. These evidences in gene data can be traced out by performing reverse engineering on the expression data. An important one of these biological phenomena is the regulatory relationship among genes.

A number of attempts have been reported in the literature toward detection of the reflection of gene regulatory network in gene expression data. Typically, the first step of such an attempt is to construct a co-expression network. There are basically two types of techniques to construct gene co-expression network (Zhang and Horvath 2005): (i) hard thresholding based and (ii) soft thresholding based. In the hard-thresholding-based scheme, two genes are connected if the expression similarity of two genes exceeds the hard threshold. Loss of information and sensitivity to the choice of the threshold are the main limitations of hard thresholding. Techniques using soft thresholding weigh each connection of the network by a number between 0 and 1.

After constructing a co-expression network, relatively dense regions in the network are extracted, which are termed as network modules. Each network module represents a regulated set of genes, which may also correspond to a set of genes associated with similar biological functions. In this article, we propose a co-expression network construction technique and a biologically relevant network module extraction technique based on a fuzzy set theoretical soft thresholding approach. The extracted network modules have been validated in terms of well-known p-value, Q-value and topological statistics using seven benchmark datasets.

Keywords. Co-expression network; fuzzy; network modules; topological property

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at: *http://www.ias.ac.in/jbiosci/jun2014/supp/Mahanta.pdf*

Published online: 1 May 2014

Table 1. Comparison of different co-expression network techniques

Method	Approach used	Measure used	Input parameters	Datasets used	Validation
FLAME	Density	Cosine correlation	1	Reduced pheripheral, blood monocytes, yeast cell cycle, mouse tissues, hypoxia response	GO Annotation
FCM	Partitional	Any	2	Yeast cell cycle	p-Value
Fuzzy-EWKM	Subspace	Any	3	Yeast cell cycle	p-Value
SYNCLUS	Partitional	Any	2	Yeast cell cycle	
Qcut	Rank-based	Pearson	0	Yeast, *Arabidopsis* and human cancer	p-Value
FUMET	Fuzzy	NMRS	2	Yeast, human and rat CNS	p-Value, Q-value and topological statistics

1.1 *Motivation*

Biologically significant genes exhibit both positive and negative correlations in gene expression data. However, based on our limited literature survey (DeSarbo *et al.* 1984; Ruan *et al.* 2010), it is observed that

(i) most existing network construction techniques are capable of handling only positive correlations;
(ii) current techniques are dependent on several input parameters and the results are highly sensitive to these parameters;
(ii) extracted network modules are validated mostly based on internal (i.e. statistical) validity measures;
(iv) hard-thresholding methods for network construction and module extraction do not preserve the continuous nature of gene expression information;
(v) effectiveness of soft-thresholding methods have already been established (Presson *et al.* 2008) for weighted gene co-expression networks in (a) preserving the gene co-expression information, (b) generating highly robust results, yet easily interpretable and (c) establishing and representing the existence of correlations of a gene in multiple classes as pointed out in Chu *et al.* (1998) and Cho *et al.* (1998).

The above limitations of the existing methods and the distinguishing features of the soft-thresholding method have motivated us to address this important research issue with the soft-thresholding method. We introduce a fuzzy-logic-based gene expression network module extraction technique that not only can identify both positively and negatively correlated co-expressed pattern but also can handle genes belonging to multiple network modules.

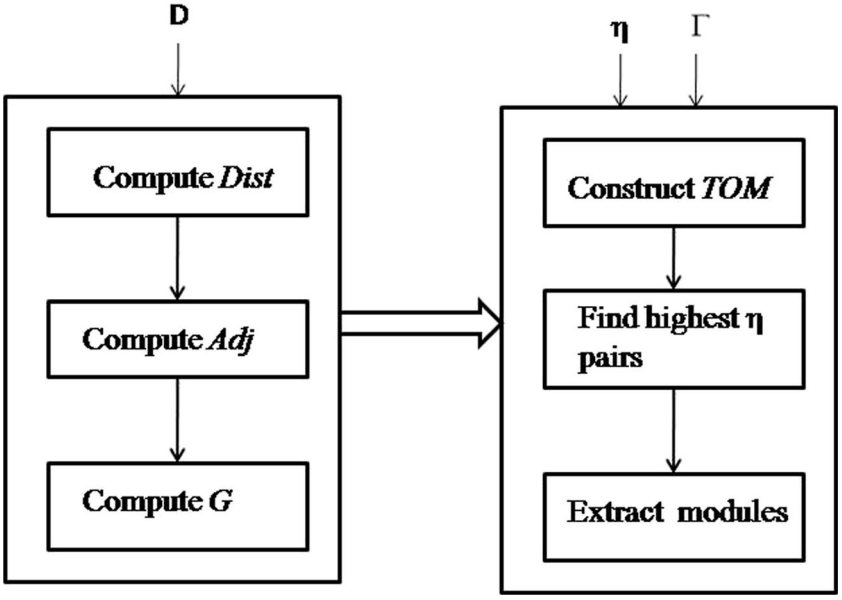

Figure 1. Methodology of proposed FUMET.

Table 2. Symbolic representations

Symbol	Meaning
G	Co-expression network
D	Gene expression data
Dist	Distance matrix
Dist(d_i, d_j)	NMRS distance between genes d_i, $d_j \in$D
Adj	Adjacency matrix
Adj(i,j)	Adjacency value between ith and jth gene
η	No of initial modules
Γ	Membership threshold
TOM(d_i, d_j)	Topological overlap value between genes d_i and d_j
TOM(C_i)	Topological overlap value of network module C_i
C	Network modules
C_i	ith network module
TOM	Topological overlap matrix

1.2 *Contributions*

The following contributions have been made in this study:

(a) A soft-thresholding co-expression network construction technique based on fuzzy logic that can handle both positive and negative correlations among genes and can handle membership of a single gene into multiple network modules
(b) Topological validation of the co-expression network
(c) A strongly correlated network module extraction technique based on fuzzy set theoretical approach
(d) Biological validation of the extracted network modules in terms of p-value, Q-value and topological properties using several benchmark gene expression datasets

1.3 *Article organization*

This article is organized as follows. In section 2, related research is reported. Our method is explained in section 3. The detailed experimental results are presented in section 4. Finally, concluding remarks and research directions are given in section 5.

2. Related work

A gene can belong to multiple functional categories in different pathways. Therefore, there may not be a clear definition of boundary between biologically relevant gene modules, and as a result, a gene may belong to multiple modules corresponding to different functional groups. Traditional methods are unable to detect the overlapped gene modules and the multifunctional behaviour of a gene. To address the above issues, soft-computing-based techniques based on variable weighing mechanism are being increasingly used by researchers. SYNCLUS (DeSarbo *et al.* 1984) is the first method for variable weighing in the traditional k-means method, which can assign a weight to each variable for the whole dataset. Besides, several other methods (Huang *et al.* 2005; Jing *et al.* 2007) have also been proposed with much flexibility in variable weighting. Fuzzy c-means (FCM) (Bezdek 1981) is another popular method that

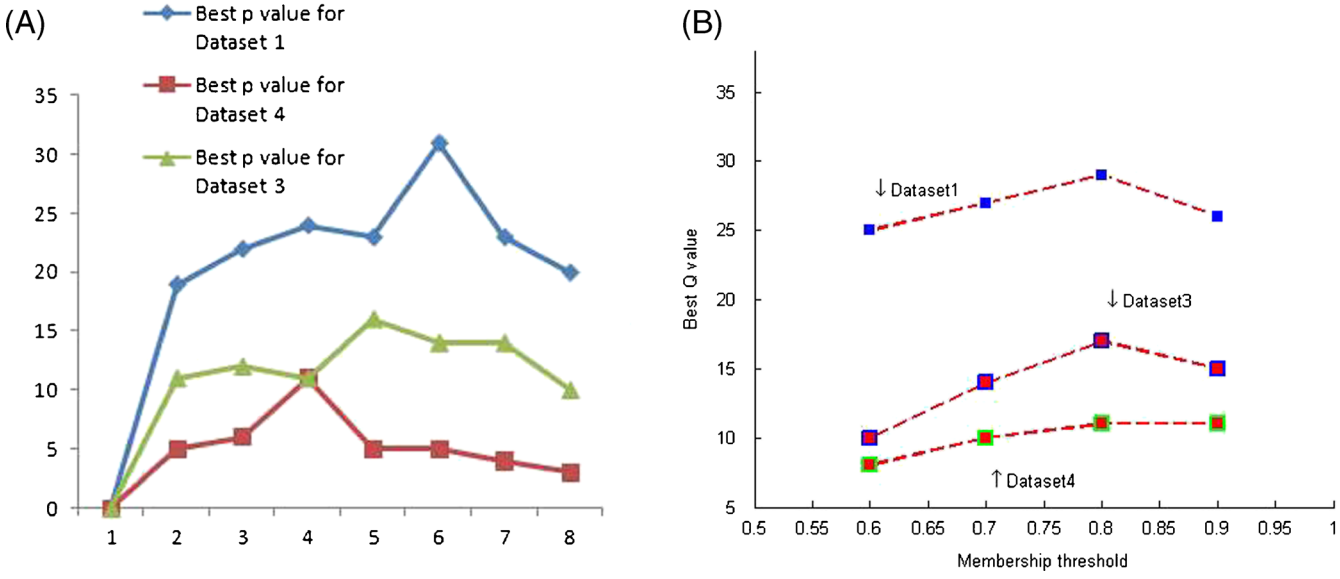

Figure 2. (**A**) X axis represents different possible values of η and Y axis represents p-values for each dataset. (**B**) Graph to determine the input parameter value Γ.

Table 3. Datasets used for evaluating FUMET

Serial no.	Dataset	No. of genes/ No. of conditions	Source
1	Yeast sporulation	474/17	*http://cmgm.standford.edu/pbrown/sporulation*
2	Human dataset	5008/4	*Sample gene in expander*
3	Subset of yeast cell cycle	384/17	*http://faculty.washington.edu/kayee/cluster*
4	Rat CNS	112/9	*http://faculty.washington.edu/kayee/cluster*
5	*Homo sapiens*	4653/30	*http://www.ncbi.nlm.nih.gov/geo*
6	Yeast dataset	698/72	*Sample gene in expander*
7	*Arabidopsis thaliana*	138/8	*http://homes.esat.kuleuven.be/sistawww/bioi/thijs/Work/Clustering.html*

provides information regarding gene multi-functionality and overlapping cellular pathways. Fuzzy-EWKM is another fuzzy-based method (Wang *et al.* 2010) which simultaneously identifies clusters of genes co-expressed under different subspaces and reveals the inter-relations between the resulting gene clusters. FLAME (Fu and Medico 2007), a variation of FCM, detects dataset-specific structures by defining neighbourhood relations among the genes, and finally neighbourhood approximation of fuzzy memberships are used to construct the modules. Many variants of FCM have been proposed with different optimizing strategies like heuristic strategies (Gasch and Eisen 2002), genetic algorithm (Hall *et al.* 1999) and expectation maximization (Nasser *et al.* 2006). Robust rough-fuzzy c-means (rRFCM) (Maji and Paul 2012) is another algorithm which integrates c-means algorithm, rough sets and probabilistic memberships of fuzzy sets while grouping functionally similar genes from microarray gene expression data. The integration of fuzzy sets enables efficient handling of overlapping partitions in noisy environment, while construction of gene co-expression network (Horvath and Dong 2008) shows that the network concept of intramodular connectivity can be interpreted as a fuzzy measure of module membership. Qcut (Ruan *et al.* 2010) consists of a rank-based network construction technique using the Pearson correlation and parameter-free module discovery technique with an objective function which optimizes modularity. A general comparison of some existing fuzzy techniques is given in table 1.

3. Methods

Our contribution mainly includes two techniques: (a) a gene co-expression network construction technique and (b) a fuzzy network module extraction technique, referred to here as FUMET. The proposed co-expression network construction technique accepts D, i.e. gene expression data, as input and constructs the network based on our proximity measure NMRS (normalized mean residue similarity) and weightage specified in the adja-

Table 4. p-Value of one of the network modules of Dataset 3

Modules	p-Value	GO number	GO category
Module1	1.53e−15	GO:0006281	DNA repair
	3.29e−14	GO:0006974	Response to DNA damage stimulus
	3.82e−14	GO:0006259	DNA metabolic process
	2.06e−14	GO:0044427	Chromosomal part
	4.81e−13	GO:0044454	Nuclear chromosome part
	1.39e−16	GO:0007049	Cell cycle
Module 2	4.16e−27	GO:0044427	Chromosomal part
	1.92e−26	GO:0006260	DNA replication
	3.19e−25	GO:0006259	DNA metabolic process
	3.64e−20	GO:0006281	DNA repair
	1.5e−20	GO:0044454	Nuclear chromosome part
	1.69e−20	GO:0006974	Response to DNA damage stimulus
	4.56e−17	GO:0051301	Cell division

cency matrix for any pair of genes (d_i, d_j). FUMET, i.e. fuzzy network module extraction technique, accepts two input parameters: number of modules (η) and membership threshold (Γ). The technique operates on the weighted co-expression network to produce network modules. A block diagram of the proposed method is given in figure 1. As can be seen from the figure, both techniques are almost balanced in terms of number of steps. Depending on the user input, it extracts number of biologically relevant highly co-expressed network modules from the co-expression network generated by the previous technique. The dependency of FUMET is discussed in a subsequent section (section 3.3).

The symbols given in table 2 and definitions reported below are useful in discussing the proposed techniques.

Definition 1 A CEN co-expression network is defined by an undirected graph G={V,E} where each v∈V corresponds to a gene and each edge e∈E corresponds to a pair of genes d_i, d_j∈D such that d_i is connected to d_j by an amount G(d_i, d_j).

Definition 2 A gene d_i is **connected** to a network module C_i, if $d_i \in C_i$ and $f_m(C_i,d_i) > \Gamma$.

Definition 3 A **network module** C_i is a set of genes forming a denser region in the co-expression network and $TOM(C_i) \in$ any top TOM values corresponding to the extracted network modules.

3.1 *Construction of co-expression network*

While constructing the co-expression network G for D, the proposed technique initially computes an adjacency matrix Adj by using a soft-thresholding technique. It encodes edge information for each pair of nodes in the co-expression network from the distance matrix. The distance matrix is obtained by using NMRS measure. The most widely used proximity measures in gene expression data analysis are Euclidean distance, Pearson correlation coefficient, Spearman correlation coefficient and mean squared residue. However, a common limitation of these measures is their incapability of handling the linear shifting patterns in the gene expression data. Mean squared residue is good enough to detect shifting patterns, but the aggregate measure cannot operate in a mutual mode, i.e. it cannot find the correlation between a pair of genes. The effectiveness of our NMRS has already been established in (Mahanta *et al.* 2012). NMRS has been found very effective in the construction of co-expression network. The NMRS of gene $d_1=(a_1, a_2,\ldots, a_n)$ with respect to gene $d_2=(b_1, b_2,\ldots, b_n)$ is defined by NMRS(d_1, d_2)=

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

where

a_{mean} is the mean of all the elements of gene d_1 and
b_{mean} is the mean of all the elements of gene d_2

The technique uses TOM (topological overlap matrix) (Zhang and Horvath 2005) to find the highest non-overlapping pairs of genes. TOM provides a similarity measure which has been found useful in biological networks. Topological overlap matrix (Zhang and Horvath 2005) is defined by

$$w_{ij}=\frac{l_{ij}+a_{\mathrm{ij}}}{\min(\mathrm{k}_i,\mathrm{k}_j)+1-a_{\mathrm{ij}}} \quad (2)$$

where $l_{ij}=\sum_u a_{iu}a_{ij}$ and $k_i=\sum_u a_{iu}$ are the node connectivity. The soft-thresholding method does not require any threshold to convert the distance matrix to adjacency matrix. The soft-thresholding function used in this study is $Adj(i,j)=Dist(d_i,d_j)$. The network construction algorithm is given in Algorithm 1.

Algorithm 1: Co-expression network construction

Input: D
Output: G

Compute *Dist* for each gene pair $d_i,d_j \in D$ using NMRS;
Compute *Adj* from *Dist* for all genes ∈ *Dist* using soft-thresholding function $Adj(i,j)=\mathrm{Dist}(d_i,d_j)$;
Construct G from *Adj* based on the association weightage, i.e. $G(d_i, d_j)$ in *Adj* against a pair of genes (d_i, d_j);

3.2 *Extraction of network modules*

FUMET accepts the co-expression network G, membership threshold and number of modules as inputs, and it extracts highly correlated network modules. The initial modules are formed with one of the gene pairs in each of these modules. Then, for each module, the membership of all the genes which are not in the module are

checked with the following membership function. The membership value of gene d_i for class C_i can be computed as

$$fm(\mathrm{C}_i, \mathrm{d}_i) = \frac{\sum_{g_j \in \mathrm{C}_i} \mathrm{Adj}_{ij}}{\min(|\mathrm{C}_i|, \mathrm{degree}(\mathrm{d}_i))} \quad (3)$$

where *degree(d_i)* is the number of nodes connected to d_i in co-expression network and Adj_{ij} is the weight of the edge corresponding to genes d_i and d_j. For a gene, if membership function produces a value greater than membership threshold against a module, the gene is included in the class. The module extraction algorithm is presented in Algorithm 2.

Algorithm 2: Network module Extraction

Input: G, η and Γ
Output: C

Compute *TOM* from *G* using equation (2);
Find η pairs corresponding to highest values in *TOM* ;
foreach $(d_i, d_j) \in$ *pairs* **do**
  Expand class C_i with initial members d_i, d_j ;
  Include $d_k \in$ G-C_i in C_i if $f_m(C_i, d_k) > \Gamma$;
  Accept C_i as a network module;
end

Lemma 1 A gene $d_i \in C_i$, may also $\in C_j$

Proof:

According to the transitive property, $\forall a,b,c \in X$: $(aRb \wedge bRc) \Rightarrow aRc$. In gene expression data, the transitive property is not followed. Let us consider two modules C_1 and C_2 which are well separated. If we consider a gene d_i similar to C_1, $d_i \in C_1$. According to transitivity property, since C_1 and C_2 are dissimilar and $d_i \in C_1$, $d_i \notin C_2$. However, since gene expression does not follow transitivity property, so C_1 may be partially similar to C_2. So $d_i \in C_2$. Again, according to definition 2, a gene $d_i \in C_i$ iff $f_m(C_i, d_i) > \Gamma$. On the other hand, it may happen that $f_m(C_j, d_i) > \Gamma$. So $d_i \in C_j$.

Hence, a gene $d_i \in C_i$ may also $\in C_j$.

3.3 *Dependency on η and Γ*

To determine the appropriate value of η and Γ, we tried an exhaustive experimentation with various possible values of and for various datasets and chose the one with highest biological significance (i.e. lowest p- or Q-value) as shown in figure 2a and b.

4. Experimental results

We implemented FUMET algorithm in MATLAB and tested it on the seven benchmark microarray datasets mentioned in table 3. 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.

4.1 *Validation*

The biological validation of the extracted network modules are performed in terms of p- and Q-value.

p-Value Biological significance of the sets of genes included in the extracted network modules are evaluated based on p-values (Tavazoie *et al.* 1999). The 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

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} \quad (4)$$

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

To compute the p-value, we used a tool called FuncAssociate (Berriz *et al.* 2003). The enriched functional categories for some of the network modules obtained by FUMET on the datasets are presented in tables 4, 5, 6 and 7. The co-expression network modules produced by FUMET contains the highly enriched cellular components of cell cycle, DNA repair (Dataset 3) , cell differentiation, cell wall assembly, spore wall assembly, developmental process, cellular component assembly involved in morphogenesis (Dataset 1), anatomical structure formation involved in morphogenesis, integral to endoplasmic reticulum membrane, intrinsic to endoplasmic reticulum

Table 5. p-Value of the network modules of Dataset 1

Modules	p-Value	GO number	GO category
Module 1	**3.38e−31**	GO:0048646	Anatomical structure formation
	5.89e−27	GO:0042244	Involved in morphogenesis spore wall assembly
	3.13e−23	GO:0048869	Cellular developmental process
	3.81e−20	GO:0070882	Cellular cell wall organization or biogenesis
	3.81e−20	GO:0071554	Cell wall organization or biogenesis
	1.06e−25	GO:0043934	sporulation
	6.10e−29	GO:0010927	Cellular component assembly involved in morphogenesis
	1.06e−25	GO:0030154	Cell differentiation
	3.53e−28	GO:0070726	Cell wall assembly
	5.89e−27	GO:0030476	Ascospore wall assembly
	5.72e−16	GO:0005628	Prospore membrane
	5.89e−27	GO:0071940	Fungal-type cell wall assembly
	2.06e−22	GO:0032502	Developmental process
	2.84e−17	GO:0048610	Cellular process involved in reproduction
	1.06e−15	GO:0006094	Gluconeogenesis
	6.60e−19	GO:0070882	Cellular cell wall organization or biogenesis
	1.43e−25	GO:0030435	Sporulation resulting in formation of a cellular spore
Module2	1.50e−15	GO:0006094	Gluconeogenesis
	6.95e−15	GO:0019319	Hexose biosynthetic process
	4.04e−19	GO:0007047	Cellular cell wall organization
	4.04e−19	GO:0045229	External encapsulating structure organization
	4.04e−19	GO:0071555	Cell wall organization
	3.61e−18	GO:0003006	Developmental process involved in reproduction
Module3	**5.13e−34**	GO:0022625	Cytosolic large ribosomal subunit
	1.1e−33	GO:0005198	Structural molecule activity
	4.74e−29	GO:0030529	Ribonucleoprotein complex
	3.64e−24	GO:0043228	Non-membrane-bounded organelle
	3.8e−26	GO:0006412	Translation
	7.12e−23	GO:0015934	Large ribosomal subunit

Modules with significantly high p-values are shown in boldface.

membrane (Dataset 2), etc., with p-values of 1.39×10^{-16}, 1.53×10^{-15}, 1.06×10^{-25}, 3.53×10^{-28}, 5.89×10^{-27}, 2.06×10^{-22}, 6.10×10^{-29}, 3.38×10^{-31} being the highly enriched ones. From the p-values given in tables 4, 5, 6 and 7, we can conclude that FUMET shows a good enrichment of functional categories and therefore projects a good biological significance.

Q-value The Q-value (Benjamini and Hochberg 1995) for a particular gene G is the proportion of false-positives among all genes which are more 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 an FDR-corrected hypergeometric test for enrichment are reported in GeneMANIA. Different GO categories of the co-expression networks produced by FUMET are displayed in tables 8, 9 and 10. The co-expression network modules produced by FUMET contains the highly enriched cellular components of regulation of transmission of nerve impulse, ascospore formation, cell wall assembly (Dataset 4), cell wall biogenesis, developmental process, ascospore wall biogenesis, anatomical structure formation involved in morphogenesis (dataset 1) etc., with Q-values of 2.50×10^{-11}, 3.68×10^{-28}, 8.73×10^{-28}, 9.46×10^{-22}, 1.46×10^{-24}, 6.20×10^{-27}, 7.48×10^{-29} being the highly enriched ones. From the results of Q-value, we arrive at the conclusion that the genes in a network module cluster obtained by FUMET seems to be involved in similar functions.

We used a Web interface GeneMANIA (Warde−Farley *et al.* 2010) to compute Q-value. Given a query list, GeneMANIA

Table 6. p-Value of the network modules of Dataset 6

Module	p-Value	GO number	GO category
Module1	2.91e−11	GO:0003723	RNA binding
	8.64e−07	GO:0006396	RNA processing
	2.17e−06	GO:0030529	Ribonucleoprotein complex
	4.51e−12	GO:0044424	Intracellular part
	7.48e−08	GO:0044464	Cell part
	3.13e−10	GO:0044444	Cytoplasmic part
	9.60e−11	GO:0008152	Metabolic process
	1.45e−10	GO:0044237	Cellular metabolic process
	1.63e−06	GO:0043170	Macromolecule metabolic process
	1.08e−06	GO:0044238	Primary metabolic process
	3.89e−08	GO:0005575	Cellular component
Module2	3.27e−20	GO:0007049	Cell cycle
	4.64e−16	GO:0044454	Nuclear chromosome part
	1.09e−16	GO:0022402	Cell cycle process
	2.72e−14	GO:0044427	Chromosomal part

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,

Table 7. p-Value of the network modules of Dataset 2

Module	p-Value	GO number	GO category
Module 1	7.51e−11	GO:0006413	Translational initiation
	1.49e−51	GO:0006412	Translation
	1.21e−10	GO:0006120	Mitochondrial electron transport, NADH to ubiquinone
	3.188e−34	GO:0005840	Ribosome
	4.88e−11	GO:0005747	Mitochondrial respiratory chain complex I
	4.88e−11	GO:0030964	NADH dehydrogenase complex
	4.88e−11	GO:0045271	Respiratory chain complex I
	1.12e−05	GO:0005689	U12-type spliceosomal complex
	2.06e−21	GO:0044449	Contractile fiber part
	4.54e−13	GO:0070469	Respiratory chain
	1.690e−05	GO:0006099	Tricarboxylic acid cycle
Module 2	2.03e−24	GO:0005743	Mitochondrial inner membrane
	1.23e−28	GO:0006091	Generation of precursor metabolites and energy
	3.26e−23	GO:0019866	Organelle inner membrane
	1.12e−23	GO:0031966	Mitochondrial membrane
	1.69e−23	GO:0055114	Oxidation-reduction process
	1.00e−19	GO:0044429	Mitochondrial part

Module with significantly high p-values is shown in boldface.

Table 8. Q-value of one of the network modules of Dataset 4

Module	GO annotation	Q-value
Module 1	Regulation of transmission of nerve impulse	2.50e−11
	Axonogenesis	2.50e−11
	Regulation of synaptic transmission	2.50e−11
	Regulation of neuron apoptosis	2.50e−11
	Axon	2.50e−11
	Regulation of neurological system process	4.20e−11
	Neuron apoptosis	4.20e−11
	Neuron death	5.33e−11
	Negative regulation of neuron apoptosis	1.56e−10
	Transmembrane receptor protein tyrosine Kinase signalling pathway	2.13e−10
	Growth factor binding	4.40e−9
	Protein autophosphorylation	1.73e−8
Module 2	Growth factor binding	4.08e−12
	Regulation of neuron apoptosis	4.08e−12
	Negative regulation of neuron apoptosis	4.91e−12
	Transmembrane receptor protein tyrosine kinase activity	1.01e−10
	Regulation of synaptic transmission	4.77e−10
	Transmembrane receptor protein kinase activity	5.58e−10
	Regulation of transmission of nerve impulse	8.06e−10
	Regulation of cell projection organization	1.25e−9
	Regulation of neurological system process	1.76e−9
	Protein tyrosine kinase activity	8.22e−9
	Protein autophosphorylation	2.78e−8

GeneMANIA also assigns coverage ratios as percentages to each of these networks with respect to the annotated genes in the genome. The percentage of co-expression on network modules produced by FUMET is given in table 11. The values are obtained by choosing the default network weighting option, i.e. the automatically selected weighing method. The network modules produced by GeneMANIA are visually presented in supplementary figure 1a–c for Datasets 4,1 and 3 respectively where the purple edge represent co-expression, light blue edges represent co-localization, green edges represent genetic interactions and dark blue represent physical interactions.

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

Module	GO annotation	Q-value
Module 1	Anatomical structure formation involved in morphogenesis	**7.48e−29**
	Sporulation resulting in formation of a cellular spore	**7.48e−29**
	Sexual sporulation	**3.68e−28**
	Ascospore formation	**3.68e−28**
	Sexual sporulation resulting in formation of a cellular spore	**3.68e−28**
	Cell development	3.68e−28
	Cellular component assembly involved in morphogenesis	6.56e−28
	Cell wall assembly	8.73e−28
	Anatomical structure development	3.61e−27
	Anatomical structure morphogenesis	3.61e−27
	Spore wall assembly	6.20e−27
	Reproductive process in single-celled organism	6.20e−27
	Ascospore wall assembly	6.20e−27
	Spore wall biogenesis	6.20e−27
	Fungal-type cell wall assembly	6.20e−27
	Ascospore wall biogenesis	6.20e−27
	Cell differentiation	1.53e−26
	Cellular developmental process	4.78e−25
	Developmental process involved in reproduction	6.37e−25
	Developmental process	1.46e−24
	Cellular component morphogenesis	1.54e−23
	Fungal-type cell wall biogenesis	1.13e−22
	Cell wall biogenesis	9.46e−22
	Sexual reproduction	7.81e−21
Module 2	Sporulation	**1.13e−29**
	Reproduction of a single-celled organism	**2.45e−21**
	Ascospore-type prospore	7.85e−18
	Intracellular immature spore	7.85e−18
	Prospore membrane	7.85e−18
	Cellular cell wall organization or biogenesis	1.069e−15
	Cell wall organization or biogenesis	1.069e−15
	Fungal-type cell wall organization or biogenesis	4.40e−15
	External encapsulating structure organization	1.02e−14
	Cellular cell wall organization	1.02e−14
	Cell wall organization	1.02e−14

Modules with significantly high Q-values are shown in boldface.

Table 10. Q-value of one of the network modules of Dataset 7

Module	GO annotation	Q-value
Module 1	Peptide transporter activity	5.74e−10
	Oligopeptide transporter activity	5.74e−10
	Peptide transport	3.48e−7
	Oligopeptide transport	3.48e−7
	Oxidoreductase activity, acting on peroxide as acceptor	2.72e−6
	Peroxidase activity	2.72e−6
	Antioxidant activity	7.54e−6
Module 2	Glutathione peroxidase activity	1.16e−17
	Oxidoreductase activity, acting on peroxide as acceptor	1.58e−10
	Peroxidase activity	1.58e−10
	Antioxidant activity	7.03e−10
Module 3	Small molecule catabolic process	1.17e−14
	Pentose-phosphate shunt	4.94e−11
	Hexose catabolic process	4.94e−11
	Monosaccharide catabolic process	4.94e−11
	Sulphate assimilation	6.64e−13

Topological validation The meaning of the edges in a gene co-expression network is a relevant question in network analysis. Different structural properties of the co-expression network can be a solution to the above question. Therefore, in this study, the structural properties of the co-expression network inferred from gene expression microarray data were compared with the topological properties of the known, well-established network data of the same organism. We use a Web application called topoGSA (Glaab *et al.* 2010) to perform a different topological validation of the extracted network modules. To reveal the meaning of the ex-tracted network modules, different network concepts that describe the network topologies in (Glaab *et al.* 2010) are used. The available network topological properties are the following:

(i) The degree of a node is the average number of edges incident to the particular node.
(ii) The shortest path length (SPL) for two nodes v_i and v_j is defined as the minimum number of edges that have to be traversed to reach v_j from v_i.
(iii) The local clustering coefficient is the probability that the neighbours of the given node are connected.
(iv) The node betweenness of a node v can be calculated from the number of shortest paths from nodes *s* to *t* going through *v*.
(v) The eigenvector centrality measures the importance of network nodes by applying a centrality score which is given by the entries of the dominant eigenvector of the network adjacency matrix.

Table 11. The weightage of co-expression by FUMET

Datasets	Network modules	Percentage
Dataset 1	C1	79.57%
	C2	88.89%
Dataset 3	C1	82.13%
	C2	88.89%
	C3	79.33%
Dataset 4	C1	78.85%

TopoGSA accepts lists of genes, proteins or microarray probe-identifiers as input, and these are mapped onto a molecular interaction network by identifying which of the molecules are present in the network. User can compare an uploaded gene/protein set against a collection of reference datasets that represent cellular pathways and processes, molecular functions or sub-cellular localizations and those that have been collected from public annotation databases including KEGG, Gene Ontology, BioCarta, InterPro and MetaCyc. This comparative analysis comes in a form of a statistics table, providing the user with the average values for the mentioned 5 topological properties computed for the uploaded gene set, 10 matched-size random gene sets in the network (as a random model) and for the entire network. From tables 14 and 15, it can be seen that the average node betweenness, degree and clustering coefficient (in some modules) of the uploaded gene set exceed the corresponding values for the matched-size random gene sets by several standard deviations and the average shortest path length were less than the corresponding value for the matched-size random gene sets. This result indicates that the genes in modules are involved in more interactions, occupy more central positions and are closer in the interaction network than random gene sets of matched sizes.

TopoGSA provides the flexibility to visualize and plot different network topological properties of genes/proteins from the uploaded dataset. A qualitative visual comparisons of the datasets median topological properties are presented in supplementary figures 2a–b and 3a–b.

The similarity ranking table of TopoGSA provides a quantitative comparison of the uploaded gene with the reference datasets, based on a similarity score. The similarity score is obtained by computing 5 ranks for each pathway/process set according to the absolute differences between each of its 5 median topological properties and the corresponding value for the uploaded gene/protein set. The sum of ranks across all topological properties is then computed and normalized to a range between 0 and 1. The network modules as shown in tables 12 and 13 extracted by FUMET show a good similarity score for two human datasets. For Dataset 2, it can be seen

Table 12. Ranking of KEGG gene sets based on topological similarity to uploaded module for Dataset 5

Identifier	Median degree	Median CC	Median SPL	Median BW	Median EVC	Score
hsa05220:Chronic myeloid leukemia	36	0.08	3.3	49218	0.1	0.93
hsa05212:Pancreatic cancer	30	0.07	3.35	45788	0.07	0.88
hsa04520:Adherens junction	26	0.05	3.38	33514	0.07	0.8
hsa04012:ErbB signaling pathway	24	0.07	3.39	28589	0.07	0.82
hsa05213:Endometrial cancer	24	0.07	3.38	33066	0.07	0.83
hsa05214:Glioma	28	0.05	3.35	37425	0.07	0.83
hsa05215:Prostate cancer	28	0.06	3.37	35660	0.07	0.84
hsa05221:Acute myeloid leukemia	27	0.09	3.39	32521	0.07	0.85
hsa05223:Non-small cell lung cancer	29	0.07	3.35	38253	0.08	0.87
hsa04320:Dorso-ventral axis formation	30	0.06	3.32	32814	0.1	0.88

Table 13. Ranking of KEGG gene sets based on topological similarity to uploaded module for Dataset 2

Identifier	Median degree	Median CC	Median SPL	Median BW	Median EVC	Score
hsa05216:Thyroid cancer	20	0.07	3.45	26256	0.05	0.83
hsa05210:Colorectal cancer	22.5	0.07	3.44	31963	0.05	0.84
hsa04520:Adherens junction	26	0.05	3.38	33514	0.07	0.85
hsa04012:ErbB signaling pathway	24	0.07	3.39	28589	0.07	0.87
hsa05214:Glioma	28	0.05	3.35	37425	0.07	0.87
hsa05213:Endometrial cancer	24	0.07	3.38	33066	0.07	0.87
hsa01510:Neurodegenerative Diseases	33	0.04	3.33	67612	0.05	0.88
hsa05215:Prostate cancer	28	0.06	3.37	35660	0.07	0.89
hsa05221:Acute myeloid leukemia	27	0.09	3.39	32521	0.07	0.9
hsa04320:Dorso-ventral axis formation	30	0.06	3.32	32814	0.1	0.91
hsa05223:Non-small cell lung cancer	29	0.07	3.35	38253	0.08	0.91
hsa05212:Pancreatic cancer	30	0.07	3.35	45788	0.07	0.92
hsa05220:Chronic myeloid leukemia	36	0.08	3.3	49218	0.1	0.96

Table 14. General topological statistics for dataset 5, a random model and the entire network

	Shortest path length	Node betweenness	Degree	Clustering coefficient	Eigen vector centrality
Uploaded Module 1	3.81	66473	23.13	0.1	0.04
10 Random simulations (mean)	4.07 (0.06)	15512 (5102)	8.99 (1.56)	0.09 (0.02)	0.02 (0)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 2	3.92	30823	14.65	0.09	0.03
10 Random simulations (mean)	4.13 (0.09)	20399 (16498)	9.27 (3.23)	0.11 (0.04)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 3	3.82	48935	19.31	0.08	0.03
10 Random simulations (mean)	4.13 (0.08)	13903 (5651)	7.78 (1.36)	0.1(0.02)	0.01 (0)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 4	3.85	40397	17.1	0.1	0.03
10 Random simulations (mean)	4.13 (0.06)	13366 (9974)	8.25 (2.79)	0.1(0.03)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)

Table 15. General topological statistics for Dataset 2, a random model and the entire network

	Shortest path length	Node betweenness	Degree	Clustering-coefficient	Eigenvector centrality
Uploaded Module 1	3.99	73124	23.64	0.1	0.04
10 Random simulations (mean)	4.06 (0.18)	15795 (14597)	9.16 (5.44)	0.01 (0.04)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 2	4	75372	24.15	0.1	0.04
10 Random simulations (mean)	4.09 (0.13)	17697 (15803)	9.67 (5.68)	0.12 (0.03)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 3	3.99	62377	21.25	0.1	0.04
10 random simulations (mean)	4.06 (0.18)	15795 (14597)	9.16 (5.44)	0.1(0.04)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)
Uploaded Module 4	3.96	69219	23.3	0.1	0.04
10 Random simulations (mean)	4.11 (0.14)	14791 (10015)	9.15 (3.65)	0.13 (0.03)	0.02 (0.01)
Static mean over entire network	4.12 (0.94)	14669 (68893)	8.27 (16.2)	0.11 (0.21)	0.02 (0.04)

Table 16. Comparison of FUMET with Qcut in terms of p-value for Dataset 3

Gene ontology ID	Gene ontology attribute	FUMET	Qcut
GO:0007049	Cell cycle	5.388e−14	1.165e−7
GO:0051301	Cell division	2.619e−7	1.547e−6
GO:0051716	Cellular response to stimulus	6.375e−8	1.872e−7
GO:0006260	DNA replication	2.124e−8	1.082e−7
GO:0044427	Chromosomal part	7.860e−10	2.274e−9
GO:0044454	Nuclear chromosome part	6.573e−9	5.784e−8
GO:0006270	DNA-dependent DNA replication initiation	2.360e−8	1.081e−7
GO:0006281	DNA repair	3.422e−9	2.970e−8
GO:0006974	Response to DNA damage stimulus	1.461e−9	2.983e−7
GO:0022402	Cell cycle process	1.482e−13	6.585e−8
GO:0048523	Negative regulation of cellular process	8.852e−9	7.970e−7
GO:0048519	Negative regulation of biological process	1.308e−8	1.073e−6
GO:0005634	Nucleus	4.693e−9	3.967e−8

from table 13 that most scores are above 0.85. Similarly, for Dataset 5, the proposed FUMET has been found capable of identifying gene modules with very high score (>0.85) tables 14 and 15.

4.2 *Comparison with Qcut and Module Miner*

Tables 16, 17, 18, 19, and 20 present some functional categories detected by FUMET which are better or equally good as compared to Qcut (Ruan *et al.* 2010) and Module Miner (Mahanta *et al.* 2012). Tables 16 and 21 show that FUMET performs better than Qcut and Module Miner in terms of

Table 17. Comparison of FUMET with Qcut in terms of Q-value for Dataset 3

GO annotation	FUMET	Qcut
Mitotic cell cycle	1.47e−19	2.41e−13
Nuclear chromosome	1.05e−14	2.41e−13
DNA repair	7.87e−14	5.25e−13
Nuclear chromosome part	1.65e−13	1.02e−12
Interphase	3.89e−13	2.45e−6
M phase	7.03e−10	2.72e−8
Cellular bud	6.22e−9	4.07e−3
Replication fork	8.852e−11	1.78e−10
Mitosis	4.12e−8	1.49e−5

Table 18. Comparison of FUMET with Qcut in terms of Q-value for Dataset 1

GO Annotation	FUMET	Qcut
Maturation of SSU-rRNA from tricistronic rRNA transcript	3.16e−9	2.07e−8
Preribosome	7.23e−41	2.44e−11
rRNA processing	2.05e−24	2.31e−6
Small-subunit processome	4.74e−11	3.73e−4
90S preribosome	2.98e−18	2.67e−12

Table 19. Comparison of FUMET with Module Miner in terms of Q-value for Dataset 1

GO Annotation	FUMET	Module Miner
Regulation of transmission of nerve impulse	2.49e−8	9.2e−7
Synapse part	7.44e−8	2.54e−5
Positive regulation of neurogenesis	1.06e−7	9.6e−5
Regulation of synaptic transmission	9.43e−10	6.43e−7
Regulation of neurological system process	1.79e−9	1.53e−6

Table 20. Comparison with Module Miner in terms of Q-value for Dataset 4

GO Annotation	FUMET	Module Miner
Regulation of synaptic transmission	2.188e−8	6.43e−7
Regulation of transmission of nerve impulse	2.49e−8	9.296e−7
Regulation of neurological system process	4.64e−8	1.53e−6
Synapse part	7.44e−8	2.54e−5
Positive regulation of neurogenesis	1.06e−7	9.60e−5

Table 21. Comparison with Module Miner in terms of p-value for Dataset 3

GO Annotation	FUMET	Module Miner
Cell cycle process	9.11e−18	1.51e−17
Nuclear chromosome part	4.64e−16	2.49e−14
Chromosomal part	6.78e−13	2.72e−14

p-value for Dataset 3. Similarly for Datasets 1, 4 and 3, FUMET performs much better than Qcut and Module Miner in terms of Q-value.

5. Conclusion and future work

In this article, a gene co-expression network construction technique based on a soft-thresholding-approach was presented. The technique has been established over seven publicly available benchmark real-life datasets. From the network, highly co-expressed modules were extracted using topological overlap matrix and a fuzzy membership function. A semi-supervised approach for construction of gene co-expression network with the support of gene ontology tools for better biological significance validated using protein interaction data is underway. Also, improvement of the existing FUMET to enable handling shifted and scaled patterns for large number of datasets is ongoing.

Acknowledgements

This article is an outcome of a research project supported by DST, India, in collaboration with CSCR, ISI, Kolkata.

References

Benjamini Y and Hochberg Y 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc*. Series B (Methodological) **57** 289–300

Berriz GF, King OD, Bryant B, Sander C and Roth FP 2003 Characterizing gene sets with FuncAssociate. *Bioinformatics* **19** 2502–2504

Bezdek JC 1981 *Pattern recognition with fuzzy objective function algorithms* (Norwell, MA, USA: Kluwer Academic Publishers)

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg T G, Gabrielian AE *et al.* 1998 A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2** 65–73

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO and Herskowitz I 1998 The transcriptional program of sporulation in budding yeast. *Science* **282** 699–705

DeSarbo WS, Carroll JD, Clark LA and Green PE 1984 Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika* **49** 57–78

Fu L and Medico E 2007 FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics* **8** 3

Gasch AP and Eisen MP 2002 Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*. **3** 0059.1–0059.22

Glaab E, Baudot A, Krasnogor N and Valencia A 2010 TopoGSA: network topological gene set analysis. *Bioinformatics* **26** 1271–1272

Hall LO, Ozyurt B and Bezdek JC 1999 Clustering with a genetically optimized approach. *IEEE Trans. Evol. Comput.* **3** 103–112

Horvath S and Dong J 2008 Geometric Interpretation of gene coexpression network analysis. *PLoS Comput. Biol*. **4** e1000117

Huang JZ, Ng MK, Rong H and Li Z 2005 Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Analysis Machine Intelligence* **27** 657–668

Jing L, Ng MK and Huang JZ 2007 An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. Knowl. Data Eng.* **19** 1026–1041

Mahanta P, Ahmed HA, Bhattacharyya DK and Kalita JK 2012 An effective method for network module extraction from microarray data. *BMC Bioinformatics* **13** S4

Maji P and Paul S 2012 Rough-fuzzy clustering for grouping functionally similar genes from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **10** 286–299

Nasser S, Alkhaldi R and Vert G 2006 A modified fuzzy k-means clustering using expectation maximization. *Fuzzy Syst*. doi: 10.1109/FUZZY.2006.1681719

Presson A, Sobel E, Papp J, Suarez C, Whistler T, Rajeevan M, Vernon S and Horvath S 2008 Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst. Biol.* **2** 95

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

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM 1999 Systematic determination of genetic network architecture. *Nat. Genet.* **22** 281–285

Wang Q, Ye Y, Huang JZ and Feng S 2010 Fuzzy soft subspace clustering method for gene co-expression network analysis; *Bioinformatics and Biomedicine Workshops (BIBMW) 2010 IEEE International Conference on* (IEEE) pp 47–50

Warde−Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, *et al* 2010 The Gen-eMANIA prediction server: biological network integration for gene prioritiza-tion and predicting gene function. *Nucleic Acids Res.* **38** W214–W220

Zhang B and Horvath S 2005 A general framework for weighted gene co-expression network analysis. *Stat. App. Genet. Mol. Biol.* **4** Article17

MS received 22 October 2012; accepted 05 February 2014

Corresponding editor: SHEKHAR C MANDE