J Biosci (2020)45:33
DOI: 10.1007/s12038-020-0007-z

X-Module: A novel fusion measure to associate co-expressed gene modules from condition-specific expression profiles

Tulika Kakati[1], Dhruba K Bhattacharyya[1]* and Jugal K Kalita[2]

[1]*Department of Computer Science and Engineering, Tezpur University, Tezpur, Assam, India*

[2]*Department of Computer Science, College of Engineering and Applied Science, University of Colorado, Colorado Springs, CO, United States*

**Corresponding author (Email, dkb@tezu.ernet.in)*

MS received 2 June 2019; accepted 7 November 2019

A gene co-expression network (CEN) is of biological interest, since co-expressed genes share common functions and biological processes or pathways. Finding relationships among modules can reveal inter-modular preservation, and similarity in transcriptome, functional, and biological behaviors among modules of the same or two different datasets. There is no method which explores the one-to-one relationships and one-to-many relationships among modules extracted from control and disease samples based on both topological and semantic similarity using both microarray and RNA seq data. In this work, we propose a novel fusion measure to detect mapping between modules from two sets of co-expressed modules extracted from control and disease stages of Alzheimer's disease (AD) and Parkinson's disease (PD) datasets. Our measure considers both topological and biological information of a module and is an estimation of four parameters, namely, semantic similarity, eigengene correlation, degree difference, and the number of common genes. We analyze the consensus modules shared between both control and disease stages in terms of their association with diseases. We also validate the close associations between human and chimpanzee modules and compare with the state-of-the-art method. Additionally, we propose two novel observations on the relationships between modules for further analysis.

Keywords. Module association; biomarkers; Parkinson's disease; Alzheimer's disease; co-expression network

1. Introduction

Co-expression of genes in a module or cluster is biologically significant. Genes in a co-expressed module are functionally related and are likely to be involved in regulatory systems (Eisen *et al.* 1998; Heyer *et al.* 1999). According to Xu et. al, genes are co-expressed across stages of a disease or process, which means that they co-regulate biological processes or functions across the stages, and therefore analysis of the potential co-regulation of co-expressed genes may be biologically significant (Xu *et al.* 2006). A CEN module extraction technique or a clustering method mines groups of genes, which are functionally coherent and exhibit similar biological patterns across different stages or conditions. In the recent past, there have been many CEN techniques and methods, which extract such co-expressed modules (Kakati *et al.* 2016; Langfelder and Horvath 2008; Leal *et al.* 2014; Mahanta *et al.* 2014; Ruan *et al.* 2010; Wang and Chen 2017). These methods extract a large number of modules at a particular stage, which may be associated with one or more modules in another stage of a disease in progression. Subsequently, a different class of methods, called differential expression analysis and differential co-expression analysis study the coherent changes of co-expressed modules and identify the genes or groups of genes which undergo coherent and significant

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12038-020-0007-z) contains supplementary material, which is available to authorized users.

Published online: 10 February 2020

changes across different stages of a disease progression or a set of conditions (Alter *et al.* 2003; Ha *et al.* 2015; Rahmatallah *et al.* 2013; Tesson *et al.* 2010; Watson 2006). However, for further analysis between a pair of modules, extracted from two different stages, there is need for a technique to identify an appropriately mapped module pair.

Langfelder and Horvath established that network methods can be used to study relationships among modules (Langfelder and Horvath 2007). They found consensus modules from two sets of modules to study relationships in common pathways and biological processes. These consensus modules were represented by eigengenes to build a network of eigengenes in each set. The correlations between eigengenes of two networks show the module relationships. These eigengene networks were used to study the relationships among modules across human and chimpanzee brains. Langfelder et al. also showed the importance of finding the preserved properties of network modules in many applications such as cholesterol biosynthesis in multiple mouse tissues, human brain and chimpanzee brain, kegg pathways between human and chimpanzee brains, and male and female cortex (Langfelder *et al.* 2011). Tan et al. found the preserved network modules using the preservation statistics described in (Langfelder *et al.* 2011), module density-based statistics, and connectivity-based statistics, and showed that there was clinical significance between preserved modules of human left atrial tissue and atrial fibrillation (Tan *et al.* 2013). Ray et al. identified co-expressed modules at an acute stage using the WGCNA framework (Ray and Bandyopadhyay 2016). For each module at the acute stage, they found preservation statistics, namely, density, cluster co-efficients, maximum adjacency ratio, intra-modular connectivity, and an eigengene based measure across chronic and non-progressor stages. Using a weighted rank aggregation scheme, the modules were ranked to assess preservation across the two phenotypes. It is an important research issue to map the modules of one stage suitably to the modules of another stage to study the relationships among co-expressed modules across a disease progression dataset. The best associated modules or clusters can also help in finding module or cluster overlap. In a gene-sample-time (GST) dataset, the overlap criterion was taken into a account to merge the clusters (biclusters) across timepoints (Zhao and Zaki 2005). The overlap measure gave module membership information and did not signify biological association between clusters. Therefore, in addition to module membership, a method needed to signify the association between module or cluster associations stages (or a set of conditions), both biologically and topologically.

To the best of our knowledge, there is no such module–module association measure or method, which can be used to assess a pair of modules undergoing associated topological and biological changes during progression of a disease. From the above discussed issues, it is understood that finding associated modules across two different stages is not a straight forward problem.

2. Methods

We propose a cross module correlation measure, referred to as X-Module to find the appropriate mapping between modules extracted from control and disease stages of two disease datasets, namely, Alzheimer's disease (AD) and Parkinson's disease (PD), and a Human-Chimpanzee (HC) dataset.

Figure 1 describes the conceptual framework of the propos1ed method. The method takes two sets of co-expressed modules extracted using a CEN technique and finds the most suitable corresponding module pairs from the sets of co-expressed modules. In this section, we discuss the steps of the proposed framework in detail.

2.1 *Data preprocessing*

In this work, we use three datasets, namely, a microarray AD dataset (GSE4226), a RNA-seq PD dataset (GSE68719), and a microarray Human-chimpanzee dataset (HC) reported by Khaitovich *et al.* (2004). The AD dataset contains 9,600 genes and is obtained from peripheral blood mononuclear cell expression of 14 elderly control and 14 AD individuals. The PD dataset consists of gene expressions for 17,580 mRNAs, collected from brain tissues of 44 control samples and 29 disease samples. For this study, we download the HC dataset from *https://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/EigengeneNetwork/* (Langfelder and Horvath 2007). The dataset was collected from brain tissues corresponding to different brain regions from three humans and three chimpanzees without any neurodegenerative disease. In the Supplementary File, we discuss the preprocessing applied to each dataset. We use the word gene(s) instead of probe(s) or mRNA(s) throughout the paper.

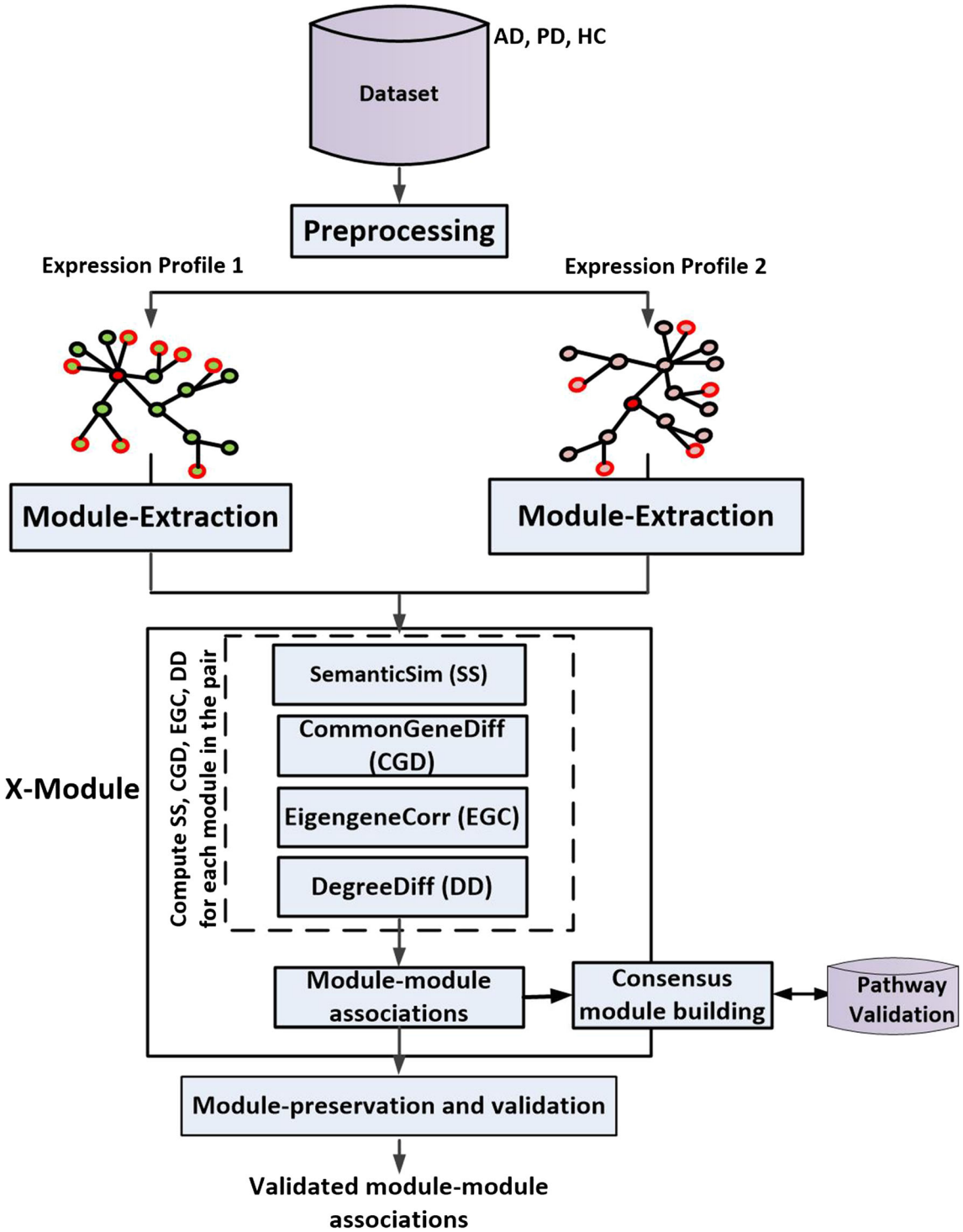

Figure 1. The X-Module framework. It finds associated modules extracted from condition-specific expression profiles. For each pair of co-expressed modules in a given dataset, we assess the topological and biological similarities in terms of their semantic similarity (SS), common gene difference (CGD), eigengene correlation (EGC), and degree difference (DD) and estimate associations. We also find the consensus modules using module-membership and eigengene correlation of the mapped modules for AD and PD datasets, and validate their significance in terms of statistical and biological values. In addition, we conduct preservation analysis of the modules using cross-tabulation of mapped module pairs extracted from AD, PD, and HC datasets, and draw a comprehensive analysis of the modules preserved in the two expression profiles of the respective datasets.

2.2 *Network module extraction*

We use THD-Module Extractor (Kakati *et al.* 2016) and WGCNA (Langfelder and Horvath 2008), respectively to extract co-expressed modules from a CEN. We find the top k biologically enriched and statistically significant modules on the basis of GO enrichment, p values, q values, degree, co-efficient, and node-betweenness. Supplementary tables 2 and 3 (see Supplementary File) show the biological and statistical significance of top k modules (inclusive of control and disease modules) extracted using THD-Module and WGCNA from AD, PD, and HC datasets, respectively.

2.3 *X-Module: module–module association*

In the recent past, many methods have been proposed to identify the changes in modules across stages or a set of conditions. These methods include differential expression analysis and differential co-expression analysis. However, there is an urgent need for a method to determine the relationships among modules across stages, or a set of conditions or time points or datasets. Preservation of modules across stages can aid in finding the appropriate mapping of modules across stages (Langfelder and Horvath 2007). X-Module finds the appropriate mapping of modules extracted from different set of conditions using three datasets. For each disease dataset, we compare the topological, and biological similarities among the mapped modules, extracted from control and disease samples. We also find the common modules, called as consensus modules using module–module membership and eigengene correlation of the mapped modules obtained from expression profiles of AD and PD datasets. These consensus modules obtained from AD and PD datasets are found to be biologically and topologically significant and are associated with pathways, which play vital role during progression of the neurodegenerative diseases. We also conduct preservation analysis of the modules using cross-tabulation of mapped module pairs extracted and draw a comprehensive analysis of the modules preserved in the two expression profiles of the respective datasets.

X-Module considers both topological and biological properties keeping in mind the fact that during progression of a disease, there are significant changes in these properties (Zhang *et al.* 2008). These properties are assessed in terms of four parameters, namely, semantic similarity, presence of common genes, correlation among eigengenes, and degree difference among the modules. The definitions used to describe our module correlation measure, the pseudo code for our algorithm X-Module Algorithm 1, and the symbols used (table 1) are discussed below.

1. *Semantic similarity between GO terms*: The genes in a module are annotated by Gene Ontology (GO) terms, and therefore the functional similarity between two modules can be defined by the semantic similarity between the GO terms annotated with the genes of the respective modules. In yesteryears, many methods were proposed to measure the semantic similarity between GO terms based on the distance from a common ancestor, (Jiang and Conrath 1997; Lin *et al.* 1998; Resnik *et al.* 1999). According to the authors in (Guo *et al.* 2006), semantic similarity measures based on information content are better than graph-structure based measures for validating the gene–gene interactions involved in a biological pathway. Therefore, the maximum similarity between the

Table 1. Symbols used in Algorithm 1 with their meanings

Symbol	Meaning	Symbol	Meaning
M^c, M^d	Set of modules in control and disease stages	Ccd	Matrix of mapped modules
W	Weight matrix of the parameters	SS_{ij}^{cd}	Semantic similarity between GO terms annotated to the gene modules i and j
$CGDg_{ij}^{cd}$	Common gene difference between modules i and j	EGC_{ij}^{cd}	Eigengene correlation of modules i and j
DDg_{ij}^{cd}	Degree difference of modules I and j	Sim_{ij}^{cd}	Similarity score of modules I and j
corrModvalue	Similarity score of the mapped module	corrModIndex	Index of the mapped module

GO terms annotated by the genes can be defined as the similarity between two genes or proteins. Among all the semantic measures and their variants, Resnik's semantic measure outperforms others in measuring gene–gene similarity. Therefore, we choose Resnik's method for mapping of modules in this article.

In our analysis, we use semantic similarity as a parameter when estimating the association between two modules. First, for each module *Mi*, we find the GO terms associated with the genes and compute the semantic similarity scores among the annotated GO terms using Resnik's semantic similarity measure. For module *Mi*, we find the average semantic similarity of the terms and denote it as *Ai*. If there are no GO terms annotating the genes of a module, then $Ai = 0$. Similarly, we find *Bj* for module *Mj* . *X* is the vector of average semantic scores for all modules at a particular stage (say, control stage) and *Y* is the vector of average semantic scores for all modules at another stage (say, disease stage). If *Si* and *Sj* are semantic similarities within modules *Mi* and *Mj,* respectively, then we define semantic similarity (SS) between the modules as follows.

$$SS(\boldsymbol{Mi}, \boldsymbol{Mj}) = 1/(abs(\boldsymbol{Si} - \boldsymbol{Sj}) + 1) \quad (1)$$

where, *Si = (Ai-min(X))/(max(X)-min(X))* and *Sj = (Bj-min(Y))/(max(Y)-min(Y))*

2. *Common gene difference*: We introduce a common gene difference (CGD) measure between two modules, Mi and Mj, which is given by the following equation.

$$CGD(\boldsymbol{Mi}, \boldsymbol{Mj}) = (|\boldsymbol{Mi} \cap \boldsymbol{Mj}|)/(|\boldsymbol{Mi} \cup \boldsymbol{Mj}|) - min(C)/(max(C) - min(C)) \quad (2)$$

where *C* represents $(|Mi \cap Mj|)/(|Mi \cup Mj|)\ \forall$ (*Mi* and*Mj)*

3. *Eigengene correlation (EGC)*: Langfelder et. al established that a module's eigengene can be used as a representative gene to summarize the module expression profiles and therefore it can be effectively used for assessing the relationship between modules (Langfelder and Horvath 2007). If Mi and Mj are two modules from control and disease stages, and matrices A and B measure the eigengene correlations between modules Mi and Mj and the expression profiles of hub gene(s) for the respective modules, then EGC for (Mi,Mj) is estimated by the following equation.

$$EGC(\boldsymbol{Mi}, \boldsymbol{Mj}) = 1/abs(max(A) - max(B) + 1) \quad (3)$$

Here, the difference between max(A) and max(B) gives the disassociation between the modules. The inverse of the difference shows the association between the modules. We add 1 to the denominator in order to avoid division by zero. Therefore, if the difference is zero (0), there is maximum association between the modules.

4. *Degree difference (DD)*: Hub gene(s) or densely connected gene(s) play an important role in many biological networks and therefore, finding hub gene(s) leads to meaningful analysis (Langfelder and Horvath 2007). In our analysis, we use the selected hub gene(s) in creating a criterion to measure the association between two modules. For a module having more than one hub gene, we find the one-hop (figure 2) neighbors of each hub gene. Suppose, the module *Mi* has more than one hub gene and *Ni* is the set of one-hop neighbors of all the hub genes of *Mi*. Similarly, we find the one-hop neighbor list *Nj* of all the hub gene(s) of *Mj* . The neighbor score for module *Mi* is $Pi = |Ni|/|Mi|$ and for module *Mj* is $Pj = |Nj|/|Mj|$ and *Pc* is the vector of neighbor scores for all modules at a particular stage (say, control stage) and Pd is the vector of neighbor scores for all modules at another stage (say, disease stage). If Di = (Pi-min(Pc))/(max(Pc)-min(Pc)) and Dj = (Pj-min(Pd))/(max(Pd)-min(Pd)), we define the DD between a pair of modules (Mi,Mj) as follows-

$$DD(\boldsymbol{Mi}, \boldsymbol{Mj}) = 1/(abs(\boldsymbol{Di} - \boldsymbol{Dj}) + 1) \quad (4)$$

Here, difference of *Di* and *Dj* shows the disassociation between the modules and inverse of the difference shows the association between the modules. Like the previous parameter, we add 1 to the denominator to prevent it from becoming 0.

The proposed method is a weight aggregation method and we assign the weights (*a*, *b*, *c*, *d*) to the parameters based on their individual significance as given in table 2. The weights are tuned to the average values of each parameter to formulate X-Module as follows.

$$X\text{-}Module(\boldsymbol{Mi}, \boldsymbol{Mj}) = a \times SS + b \times CGD + c \times EGC + d \times DD \quad (5)$$

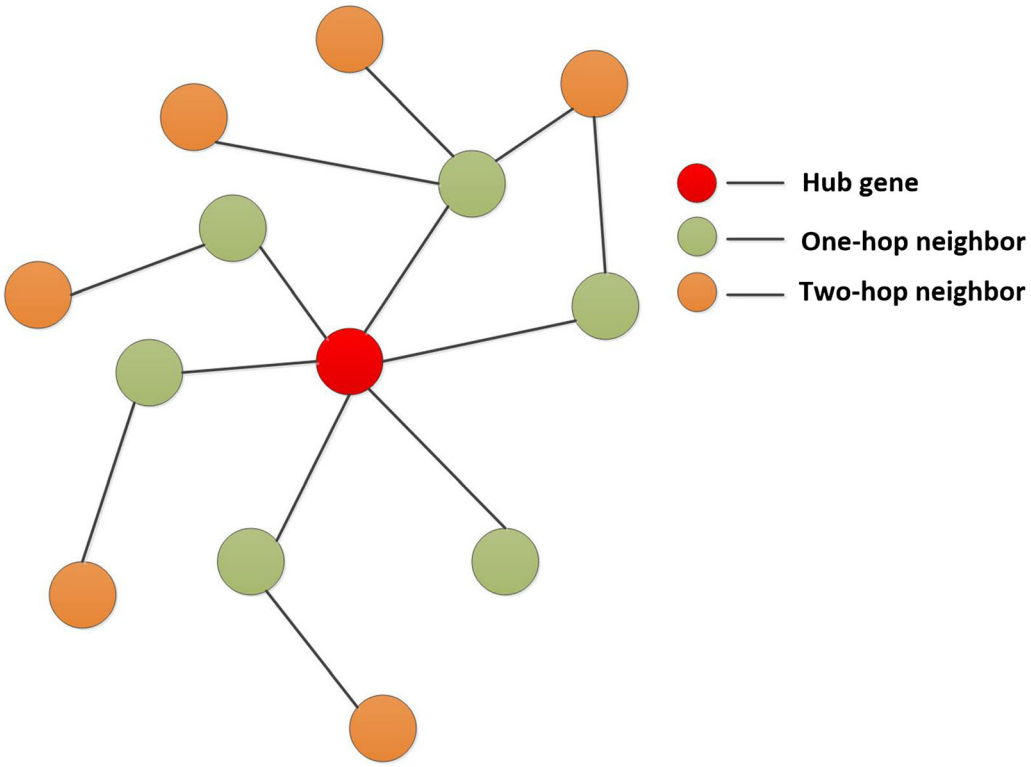

Figure 2. A module with hub gene (red colored node), one-hop neighbors (green colored nodes), and two-hop neighbors (orange colored nodes).

Algorithm 1: X-Module

Input: M^c={M_1^c, M_2^c , M_3^c, ……,M_m^c}, M^d={M_1^d, M_2^d , M_3^d, ……,M_n^d}, W ={w_1, w_2, w_3, w_4 }, m,n= number of control and disease modules, respectively.

Output: C^{cd}, where $C_{i1}^{cd} \triangleq C_{i2}^{cd}$ means C_{i1}^{cd} corresponds to C_{i2}^{cd}. Here, i=1, 2, . . . , m and j=1, 2

Initialization: C^{ij}=φ

foreach *(i in m)* **do**

  foreach *(j in n)* **do**

    $SS_{ij}^{cd} = 1/(S_i^c - S_j^d)$;

    $CGDg_{ij}^{cd}$=cardinal (∀ x intersect (M_{ix1}^c , M_{jx1}^d));

    $EGC_{ij}^{cd} = corr(E_i^c, E_j^d)$;

    $DD_{ij}^{cd} = 1/(D_i^c - D_j^d)$;

    call Decision($SS_{ij}^{cd}, CGDg_{ij}^{cd}, EGC_{ij}^{cd}$, DD_{ij}^{cd}, i, j, W);

    X-Module (M_i^c, M_j^d) = w_1 x SS_{ij}^{cd} + w_2 x $CGDg_{ij}^{cd}$ + w_3 x EGC_{ij}^{cd} + w_4x DD_{ij}^{cd};

  — Sim_{ij}^{cd} = X-Module(M_i^c, M_j^d)

—

foreach *(i in m)* **do**

  [corrModValue corrModIndex] = Sim_i^{cd};

  j=1;

  $C_{ij}^{cd} \leftarrow$ i;

  j=2;

  $C_{ij}^{cd} \leftarrow$ corrModeIndex;

—

return C^{cd}

Table 2. Decision (SS_{ij}^{cd})

SS_{ij}^{cd}	$CGDg_{ij}^{cd}$	EGC_{ij}^{cd}	DD_{ij}^{cd}	Decision
1	1	1	1	1
1	1	1	0	1
1	1	0	1	1
1	0	1	1	1
0	1	1	1	1
1	1	0	0	1
1	0	1	0	1
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	0	0	0
0	0	0	1	0
0	1	0	0	0
0	0	1	0	0
0	0	0	0	0
0	0	1	1	0

2.4 *Finding consensus modules*

Study of relationships among modules across groups of conditions or datasets is interesting because these relationships capture the biological changes during progression of a disease (Langfelder and Horvath 2007). Consensus modules are defined as groups of genes correlated across conditions in a dataset. To find consensus modules across control and disease conditions of AD and PD datasets, we consider two parameters, namely, module membership and eigengene correlation. The definitions of module membership and eigengene correlation and pseudocodes for finding consensus modules are given in the Supplementary file.

2.5 *Preservation of the mapped modules*

X-Module finds the association between a pair of modules across control and disease conditions of AD and PD datasets based on estimates of four parameters as discussed above. To the best of our knowledge, there is no tool that can be used to validate the correctness of associations given by X-Module. Langfelder et. al used differential eigengene network analysis to assess network properties acrosss CEN modules. In this work, we first use standard cross-tabulation to show the appropriate mapping of the corresponding module pairs across control and disease stages. We also adapt network-based preservation statistics, namely, network density, connectivity, and separability, defined by Langfelder et. al to ascertain the amount of module preservation among the mapped modules across control and disease states. Network density assesses whether the mapped modules remain densely connected across control and disease conditions or datasets. Using connectivity, the hub gene(s) of the mapped modules are examined to find whether they are preserved across control and disease sets of conditions. Separability of the mapped modules measures the correlations and distinctness of the modules across control and disease conditions. These network statistics are summarized together to formulate *ZSummary* (Langfelder and Horvath 2007) of the modules obtained from disease stage (test modules) across control stage (reference modules).

3. Results

In this section, we discuss the results on AD, PD, and HC datasets for the co-expressed modules extracted using two CEN module extraction methods, THD-Module Extractor and WGCNA and show the generalizability of X-Module. X-Module is independent of the module extraction approach used in any CEN method and is capable of finding module-module associations between co-expressed modules pairs extracted from expression profiles of two different stages.

3.1 *Analysis of co-expressed modules obtained using THD-Module Extractor and WGCNA*

X-Module accepts two sets of co-expressed modules to find the most correlated modules across the sets. We analyze top k control and disease modules obtained using THD-Module Extractor and WGCNA in terms of topological and biological significance. From tables 2 and 3 of Supplementary File, it is evident that the co-expressed modules are highly enriched with GO attributes with low p and q values. In addition, the values of the network statistics defining the topological properties of the co-expressed modules such as degree, co-efficient, and node-betweenness are high.

3.2 *Analysis of consensus modules across groups of samples of datasets*

From AD expression profiles, we find 8 consensus modules. These modules share common genes and have higher correlation among the module eigengenes.

In table 3, we describe the top k=5 consensus modules obtained using THD-Module Extractor with the number of common genes, eigengene correlations, p, q,

Table 3. Biological and statistical analysis of consensus modules of AD obtained using THD-Module Extractor

Module	Number of common genes	Eigengene correlation	Co-efficient	Node-betweenness	Degree	*p*	*q*	Significant pathways
1	140	0.810662094	11.86	0.12	29514	1.048E−5	3.717E−2	Apoptosis signaling pathway JAK/STAT signaling pathway Wnt signaling pathway p53 pathway Alzheimer's disease-presenilin pathway Parkinson's disease
2	74	0.998976033	16.33	0.13	47065	3.658E−6	7.763E−3	Integrin signalling pathway Alzheimer's disease-presenilin pathway p53 pathway Dopamine receptor mediated signaling Pathway Parkinson's disease Adrenaline and noradrenaline biosynthesis
3	67	0.978293204	14.63	0.07	38742	1.963E−5	4.149E−2	Apoptosis signaling pathway JAK/STAT signaling pathway Adrenaline and noradrenaline biosynthesis Integrin signalling pathway Parkinson's disease Inflammation mediated by chemokine and cytokine signaling pathway
4	61	0.978142183	8.31	0.09	16469	1.059E−5	7.308E−3	Apoptosis signaling pathway Adrenaline and noradrenaline biosynthesis Parkinson's disease Nicotinic acetylcholine receptor signaling Pathway Inflammation mediated by chemokine and cytokine signaling pathway Interleukin signaling pathway
5	71	0.842823263	8.68	0.13	17131	3.120E−6	9.265E−4	Apoptosis signaling pathway Alzheimer's disease-presenilin pathway p53 pathway Dopamine receptor mediated signaling Pathway Inflammation mediated by chemokine and cytokine signaling pathway Interleukin signaling pathway

degree, co-efficient, node-betweenness, and pathways associated with AD. Pathway analysis provides better understanding of the biological significance of the co-expressed modules. Pathways namely Apoptosis signaling, p53 pathway, Parkinson disease, Alzheimer disease-presenilin, Inflammation mediated by chemokine and cytokine signaling pathway are shared by the consensus modules. Moreover, the p and q values of the consensus modules are lower, which signify the statistical significance of the modules. For example, the consensus module 1 consists of 140 genes shared among control and disease stages with eigengene correlation of 0.810662094. In addition, pathways such as Apoptosis signaling pathway, Wnt signaling pathway, Alzheimer disease-presenilin pathway shared by the genes of consensus module 1 are associated to AD (De Ferrari and Inestrosa 2000; Thinakaran 1999). Thus, from this analysis, we can draw the conclusion that the consensus modules are both biologically and statistically significant.

3.3 *Module preservation of the mapped module pairs for AD and PD*

We show the cross-tabulation of the mapped module pairs using the number of common genes and p values of each module pair. For example, from table 4, it is evident that the modules 4 and 7 extracted using THD-Module Extractor show close association across control and disease stages of the AD dataset with 122 common genes with p value of 1.58e−07. Similarly, in table 5, we see that module 3 of control stage and module 5 of disease stage extracted using THD-Module Extractor have 72 common genes shared by them with p value of 1.68E−13. Again, in table 6, we find that for modules extracted using WGCNA, the Chimpanzee modules matches well in terms of p values and common genes with the human modules. Thus, we can demonstrate the modules across control and disease stages share overlapping genes with low p values. However, there are some serious disadvantages of the standard cross-tabulation based statistics. Cross-tabulation is defined for modules obtained from only clustering is based on overlap criterion and is not suitable for determining module preservation. In (Langfelder and Horvath 2007), the authors introduced network statistics, namely network density, connectivity, and separability of the modules to measure network preservation using *modulePreservation()*. For example, figure 3 shows that the median rank and *Zsummary* of the modules obtained from control and disease stages of AD dataset are higher than the significance level. However, we did not consider the preservation of biological functionality of the modules. Table 7 shows that for the modules obtained using WGCNA, pathway significance values of the mapped modules obtained using X-Module are better in terms of p values than the preserved modules obtained using *modulePreservation()*. For example, the biological pathways, such as Apoptosis signaling pathway, Alzheimer disease-amyloid secretase pathway, Wnt signaling pathway, and Parkinson disease

Table 4. Standard cross-tabulation of mapped modules extracted using THD-Module Extractor from AD dataset

Control module	Disease module	p value	Common number of genes
Module 1	Module 1	1.83E−05	140
Module 2	Module 2	1.13E−07	155
Module 3	Module 2	2.35E−08	58
Module 4	Module 7	1.58E−07	122
Module 5	Module 5	7.45E−09	90
Module 6	Module 5	7.45E−07	23
Module 7	Module 7	7.45E−07	71
Module 8	Module 1	1.36E−06	71

Table 5. Standard cross-tabulation of mapped modules extracted using THD-Module Extractor from PD dataset

Control module	Disease module	p value	Common number of genes
Module 1	Module 5	1.68E−13	63
Module 2	Module 2	2.53E−13	76
Module 3	Module 5	1.68E−13	72
Module 4	Module 5	1.68E−13	62
Module 5	Module 6	3.08E−13	78
Module 6	Module 8	2.64E−13	50
Module 7	Module 8	2.64E−13	80
Module 8	Module 1	2.53E−13	70

Table 6. Standard cross-tabulation of mapped modules extracted using WGCNA from HC dataset

Control module	Disease module	p value	Common number of genes
Module 1	Module 3	1.46E−07	185
Module 2	Module 2	1.44E−07	96
Module 3	Module 2	1.18E−08	95
Module 4	Module 5	1.46E−07	85
Module 5	Module 5	1.21E−07	44
Module 6	Module 1	1.15E−07	68

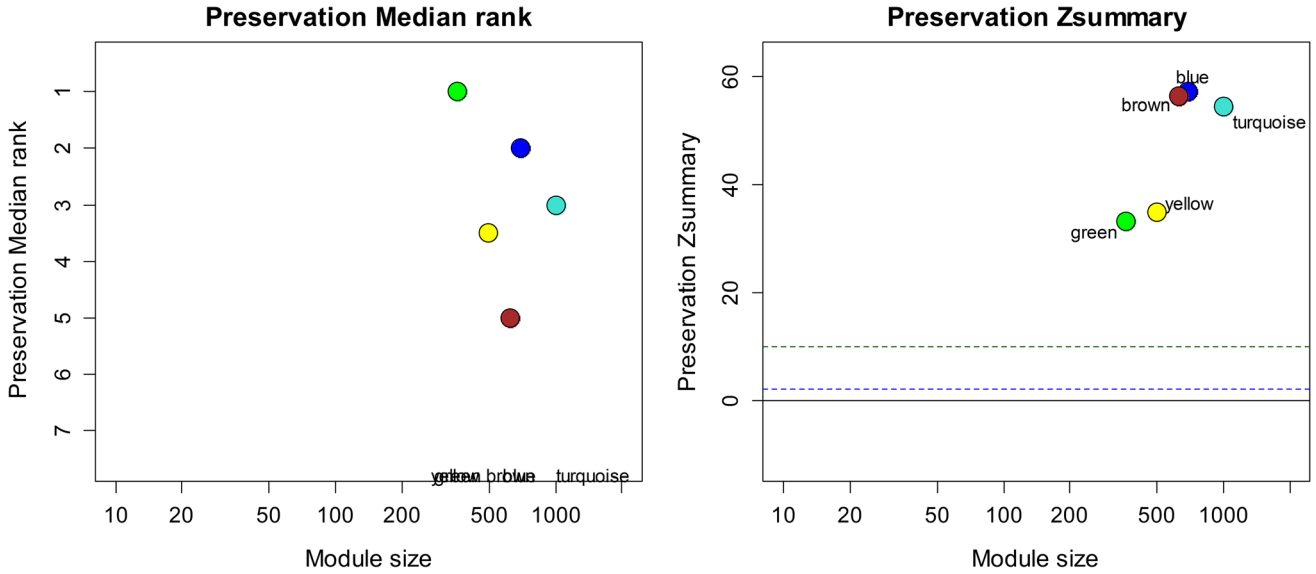

Figure 3. Median rank preservation and *Zsummary* preservation of AD disease (test) modules across control (reference) modules. Here, the median rank and *Zsummary* of all the modules are above the significant threshold and therefore, the modules are preserved across control and disease expression profiles of AD.

pathway shared by the mapped modules obtained using X-Module show lower *p* values in comparison with the pathways regulated by the preserved modules obtained using *modulePreservation()*.

3.4 *X-Module package*

We have developed an open source R package, called X-Module for better understanding and to facilitate the users to enable interested readers to reproduce the results. X-Module finds the most apposite module pairs using both microarray and RNA-seq count data. The package includes all the necessary scripts and datasets for running a demonstrative example and is available at *https://github.com/tulikakakati/X-Module*.

4. Discussion

In this paper, we highlight the important issue of analysis of stage or condition-specific modules after extraction of co-expressed modules across control and disease stages. In the recent past, there has been ample research on the study of significant changes across stages, of the co-expressed modules extracted using module extraction techniques. However, in order to assess a pair of co-expressed modules and find their preservation properties, we need to find the corresponding pair of modules across control and disease stages of a disease. There is limited amount of work on the study of relationships of modules across different stages of a disease or multiple conditions. X-Module finds the corresponding module pairs from expression profiles obtained from AD, PD, and HC datasets. At first, the co-expressed modules are extracted for control and disease stages of AD and PD dataset using module extraction techniques, i.e,. WGCNA and THD-Module Extractor. The corresponding module pairs, obtained using X-Module are then statistically and biologically validated in terms of GO enrichment and statistical enrichments. We also find biologically meaningful pathways associated with AD and PD from the consensus modules obtained from AD and PD datasets using pathway analysis. Pathways such as Apoptosis signaling pathway, Wnt signaling pathway, and Alzheimer's disease-presenilin pathway shared by the genes of consensus module 1 are associated with AD (De Ferrari and Inestrosa 2000; Thinakaran 1999).

The appropriate module pairs obtained from AD, PD, and HC datasets using X-Module are validated using standard cross-tabulation in terms of common genes and eigengene correlations.

We believe that X-Module will benefit the users in finding close associations between co-expressed modules, extracted from two different stages in terms of commonality of genes, semantic similarity, eigenegene correlation and their degrees of connectivity. In other words, if two modules are associated and are validated using X-Module, then the two modules have the following features:

Table 7. Comparison of PD preserved modules in terms of preservation statistics (*P restatistics*) and biological statistics (Pathway significance) obtained from X-Module and *modulePreservation()* using WGCNA

Method	Module pair	*Prestatistics*	Pathway significance		
			Pathways	*p* value	*q* value
X-Module	(*M*1,*M*1)	3.358233254	Apoptosis signaling pathway	2.864E−18	7.255E−15
	(*M*2,*M*4)	2.842503479	VEGF signaling pathway	1.1211E−10	2.012E−8
	(*M*3,*M*4)	2.44366501	Alzheimer disease-amyloid secretase pathway	3.12E−14	4.07E−11
	(*M*4,*M*6)	3.082228178	Apoptosis signaling pathway	1.01E−11	2.18E−9
	(*M*5,*M*2)	2.956770598	Alzheimer disease-presenilin pathway	3.01E−17	2.09E−12
	(*M*6,*M*4)	2.779362029	Wnt signaling pathway	2.01E−15	5.11E−12
	(*M*7,*M*7)	2.592322048	p53 pathway feedback loops 2	1.15E−14	1.19E−11
	(*M*8,*M*5)	2.932836494	Parkinson disease	2.16E−15	3.14E−12
modulePreservation()	turquoise	42	Apoptosis signaling pathway	1.604E−08	1.513E−05
	blue	20	Wnt signaling pathway	7.11E−05	9.21E−02
	pink	15	Apoptosis signaling pathway	5.78E−08	7.45E−03
	yellow	10	Parkinson disease	1.08E−7	5.91E−05
	green	8	p53 pathway feedback loops 2	5.45E−08	5.06E−05
	black	7	Wnt signaling pathway	7.22E−05	2.61E−04
	brown	1.5	Parkinson disease	3.23E−08	2.24E−06

1. Common genes between the associated modules signify the co-expression of the common genes across different stages (biological or functional).
2. Maximum semantic similarity signifies that the two associated modules will have genes showing maximum semantic similarity (biological or functional).
3. Associated modules have maximum correlations between their eigengenes (topological).
4. Associated modules have maximum connectivity to the hub genes (topological and biological).

Thus, the mapped modules have both biological and topological associations between them across different stages, which provide insights into the common behavioural changes of the apposite modules during progression of any diseases.

Finally, we present the following two properties of the X-Module measure.

P1 Non-negativity: If M_i^{s1} and M_j^{s2} are two modules of states s1 and s2 and $M_i^{s1} \triangleq M_j^{s2}$, then X-Module(Mi,Mj)$\geq$0.

Explanation: X-Module measures the appropriate associations between a pair of modules in terms of the weighted aggregation of four parameters, namely, SS, CGD, EGC, and DD. From equations 1, 2, 3, and 4, we can estimate that the values of SS, CGD, EGC, and DD are always greater or equal to zero (0). Moreover, the weight aggregates a, b, c, and d is the means of SS, CGD, EGC, and DD, and are always positive. Therefore, from equation 1, we can conclude that the X-Module (Mi,Mj) $\geq$ 0.

P2 Symmetricity: If M_i^{s1} and M_j^{s2} are two modules of states s1 and s2 and $M_i^{s1} \triangleq M_j^{s2}$, then $M_j^{s2} \triangleq M_i^{s1}$.

Explanation: Two modules M_i^{s1} and M_j^{s2} from two different stages s1 and s2 are found to have close association to each other if and only if they share a large number of common genes Cg_{ij}^{12} with higher semantic similarity S_{ij}^{12}, eigengene module correlations E_{ij}^{12}, and low degree difference D_{ij}^{12}. Since, it is understood that for two modules M_i^{s1} and M_j^{s2}, $Cg_{ij}^{12} = Cg_{ij}^{21}$, $S_{ij}^{12} = S_{ij}^{21}$, $E_{ij}^{12} = E_{ij}^{21}$, $D_{ij}^{12} = D_{ij}^{21}$ and the correspondence between two modules depends on sum of these four variables (from Algorithm X-Module given in the Supplementary File); therefore, it is established that if $M_i^{s1} \triangleq M_j^{s2}$, then $M_j^{s2} \triangleq M_i^{s1}$.

P3 Transitivity: The following conditions are not necessary and sufficient for two modules of two distinct stages, e.g., M_i^{s1} and M_k^{s3} to correspond to each other:

(i) $M_i^{s1} \triangleq M_j^{s2}$ and

(ii) $M_j^{s2} \triangleq M_k^{s3}$, where M_j^{s2} is a module from the intermediate stage.

Explanation: Let us prove the statement by contradiction. Let us assume that M_i^{s1} and M_j^{s2} depend on the sum of Cg_{ij}^{12}, S_{ij}^{12}, E_{ij}^{12}, and D_{ij}^{12}, where Cg_{ij}^{12} is the

common gene difference between modules M_i^{s1} and M_j^{s2}. Assume that $M_i^{s1} \triangleq M_j^{s2}$ and $M_j^{s2} \triangleq M_k^{s3}$ and also $M_i^{s1} \triangleq M_k^{s3}$.

But if M_i^{s1} = {g1, g2, g3, g4} and M_j^{s2} = {g2, g3, g5, g6}, then

$M_i^{s1} \bigcap M_j^{s2}$ = {g2, g3} and $Cg_{ij}^{12} = 2$

Similarly, if M_j^{s2} = {g2, g3, g5, g6} and M_k^{s3}={g5, g6} then,

$M_j^{s2} \bigcap M_k^{s3}$ = {g5, g6} and $Cg_{ij}^{12} = 2$

Here, $M_i^{s1} \bigcap M_k^{s3} = \varphi$ and $Cg_{ik}^{13} = 0$

Therefore, we have a contradiction and so even if $M_i^{s1} \triangleq M_j^{s2}$ and $M_j^{s2} \triangleq M_k^{s3}$, $M_i^{s1} \triangleq M_k^{s3}$.

The proposed method can help researchers find the associated module pairs across normal and disease stages.

During analysis of the experimental results, we witness two scenarios.

1. We observe a one-to-many relationship where a module in control stage may map to more than one module in disease stage. This scenario is justified because the participating genes of a co-expressed module have the potential to co-regulate multiple biological functions (Lobo 2008).
2. We also observe a many-to-one relationship where more than one module corresponds to a single module from the disease stage. In support of this scenario, we suggest that this may occur because multiple genes or modules regulate same biological functions (Eisen *et al.* 1998; Heyer *et al.* 1999).

There is no standard validation tool to evaluate the accuracy of the proposed method. We evaluate the accuracy of our method by extending our analysis to pathway enrichment and GO enrichment analysis of the module pairs obtained by WGCNA and THD-Module Extractor. From pathway analysis, we observe that most of the pathways shared by the module pairs are enriched equally in terms of both p and q values. Similarly, the GO terms annotated by the genes of module pairs are enriched significantly in terms of p and q values. Work is ongoing to extend the X-Module framework to consider module association finding for one species against a large number of species across the states using py-CUDA platform. Further, disease biomarker identification for a given disease as well as for a group of diseases is also underway to support subsequent analysis of both microarray and RNASeq data using a distributed computing framework.

References

Alter O, Brown PO and Botstein D 2003 Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Nat. Acad. Sci*. **100** 3351–3356

De Ferrari GV and Inestrosa NC 2000 Wnt signaling function in Alzheimer's disease. *Brain Res. Rev.* **33** 1–12

Eisen MB, Spellman PT, Brown PO and Botstein D 1998 Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci*. **95** 14863–14868

Guo X, Liu R, Shriver CD, Hu H and Liebman MN 2006 Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* **22** 967–973

Ha MJ, Baladandayuthapani V and Do K-A. 2015 Dingo: differential network analysis in genomics. *Bioinformatics* **31** 3413–3420

Heyer LJ, Kruglyak S and Yooseph S 1999 Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **9** 1106–1115

Jiang JJ and Conrath DW 1997 Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008

Kakati T, Kashyap H and Bhattacharyya DK 2016 Thd-module extractor: an application for cen module extraction and interesting gene identification for Alzheimer's disease. *Sci. Rep*. **6** 38046

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do H-H, Weiss G, Enard W, *et al.* 2004 Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14** 1462–1473

Langfelder P and Horvath S 2007 Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1** 54

Langfelder P and Horvath S 2008 WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9** 559

Langfelder P, Luo R, Oldham MC and Horvath S 2011 Is my network module preserved and reproducible? *PLoS Comput. Biol*. **7** e1001057

Leal LG, López C and López-Kleine L 2014 Construction and comparison of gene co-expression networks shows complex plant immune responses. *PeerJ* **2** e610

Lin D *et al.* 1998 An information-theoretic definition of similarity. *ICML* **98** 296–304

Lobo I 2008 Pleiotropy: one gene can affect multiple traits. *Nat. Edu*. **1** 10

Mahanta P, Ahmed HA, Bhattacharyya DK and Ghosh A 2014 Fumet: a fuzzy network module extraction technique for gene expression data. *J. Biosci*. **39** 351–364

Rahmatallah Y, Emmert-Streib F and Glazko G 2013 Gene sets net correlations analysis (GSNCA): a multi-variate differential coexpression test for gene sets. *Bioinformatics* **30** 360–368

Ray S and Bandyopadhyay S 2016 Discovering condition specific topological pattern changes in coexpression network: an application to HIV-1 progression. *IEEE/ACM Transact. Comput. Biol. Bioinform*. **13** 1086–1099

Resnik P *et al*. 1999 Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11** 95–130

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

Tan N, Chung MK, Smith JD, Hsu J, Serre D, Newton DW, Castel L, Soltesz E, Pettersson G, Gillinov AM, *et al.* 2013 A weighted gene co-expression network analysis of human left atrial tissue identifies gene modules associated with atrial fibrillation. *Circ. Genomic Precision Med*. **6** 113

Tesson BM, Breitling R and Jansen RC 2010 Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform*. **11** 497

Thinakaran G 1999 The role of presenilins in Alzheimer's disease. *J. Clin. Invest.* **104** 1321–1327

Wang Q and Chen G 2017 Fuzzy soft subspace clustering method for gene co-expression network analysis. *Int. J. Machine Learn. Cybernetics* **8** 1157–1165

Watson M 2006 Coxpress: differential co-expression in gene expression data. *BMC Bioinformatics* **7** 509

Xu X, Lu Y, Tung A and Wang W 2006 Mining shifting-and-scaling co-regulation patterns on gene expression profiles. *Proceedings of the 22nd International Conference on Data Engineering ICDE'06* pp 89–89 (IEEE)

Zhang B, Li H, Riggins RB, Zhan M, Xuan, J., Zhang, Z., Hoffman EP, Clarke R and Wang Y 2008 Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics* **25** 526–532

Zhao L and Zaki MJ 2005 Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. *Proceedings of the 2005 ACM SIGMOD international conference on Management of Data* pp 694–705 (ACM)

Corresponding editor: Stuart A Newman