MODA: MOdule Differential Analysis for weighted gene co-expression network

Dong Li
School of Computer Science, The University of Birmingham, UK
James B. Brown
Department of Statistics, University of California Berkeley, USA
Luisa Orsini
School of Biosciences, The University of Birmingham, UK
Zhisong Pan,Guyu Hu
PLA University of Science and Technology, China
Shan He
School of Computer Science, The University of Birmingham, UK

May 18, 2016

1 Summary

Gene co-expression network differential analysis is designed to help biologists understand gene expression patterns under different condition. By comparing different gene co-expression networks we may find conserved part as well as condition specific set of genes. Taking the network as a collection as modules, we use a sample-saving method to construct condition-specific gene co-expression network, and identify differentially expressed subnetworks as conserved or condition specific modules which may be associated with biological processes. We have implemented the method as an R package which establishes a pipeline from expression profile to biological explanations. The usefulness of the method is also demonstrated by synthetic data as well as Daphnia magna gene expression data under different environmental stresses.
Availability: Available at https://www.cs.bham.ac.uk/ szh/software.xhtml
Contact: s.he@cs.bham.ac.uk

2 Introduction

Gene co-expression network attracts much attention nowadays. In such a network, nodes represent genes and each edge connecting two genes stands for how much degree may this pair of genes are co-expressed across several samples. The presence of these edges is commonly based on the correlation coefficients between each gene pairs. The higher of correlation between a pair of genes, the higher probability that there exists a co-functionality relationship between them. With proper choice of minimal correlation value as a threshold, we can generate an unweighted and undirected network for given gene expression profile. But the optimal cut-off threshold is difficult to determine. And throwing away relatively large proportion of correlation coefficients will lead to information loss. In contrast, weighted correlation network analysis (WGCNA) overcomes this drawback by keeping all possible edges but shows how significant is the co-expression relationship using edge weights [1, 2].

A module in a biological network is defined as a subnetwork which may involves a common function in biological processes. The module detection in WGCNA is based on hierarchical clustering, which groups similar genes into one cluster. The similarity was defined by topological overlap measure [2]. Following the logic of WGCNA, here we mainly improve it from the following three aspects: 1) How to determine the cutting height of hierarchical clustering tree roughly depends on self-definition in WGCNA. Here we give an option to choose the height based on the quality of partition. 2) Edge weights in gene co-expression networks are defined by correlation coefficients of gene pairs. And it is well known that the accurate correlation coefficient is approximated by $1/sqrt(n)$ where n is the number of samples, which makes it impossible to get reliable correlation coefficients with only several replicates under each experimental condition in practice. We use a sample-saving way to analyze condition-specific co-expression network for each single condition. 3) Taking a network as a collection of modules, we generalize the differential analysis from individual genes to modules, which may find condition specific and conserved subnetworks.

3 Methods

Inspired by the concept of partition density of link communities [3] where the modules were defined based on the link similarity, we propose a cutting method to make the average density of resulting modules to be maximal. Here we simply define the module density as the average edge weights in one module (equation (1) in supplementary file) which keeps the same in [2], and then find the cutting height of

hierarchical clustering that leads to maximal average density. We also provide other criterion such as average modularity for weighted network [4] of resulting clusters to determine the cutting height.

General gene differential analysis has covered identification of important individual genes which shows significant changes across multiple conditions [5]. However, based on the fact that genes interact with each other to exert some biological function instead of acting alone, it is more informative to identify a subnetwork (module) of genes which are conserved across multiple conditions or just active in certain conditions. DICER [6] also goes beyond individual gene differential analysis, using a probabilistic framework to detect differentially co-expressed gene sets. DINA [7] can identify condition-specific modules from a collection of condition-specific gene expression profiles which differs from our sample-saving method. Based on a set of condition-specific networks, we use WGCNA to identify modules for different networks. Then, we use the Jaccard index, which essentially measures the similarity between two sets of elements, to measure the similarity between modules from two different networks.

By comparing all module pairs of two networks, we can get a similarity matrix A,where each entry A_{ij} means the Jaccard similarity coefficient between the i-th module from the network N_1 and j-th module from the network N_2. Assume the N_1 is background, normally containing samples from all conditions, and the N_2 is constructed from all samples minus samples belong to certain condition D [8]. Then the elements in row sum of A (vector denoted by $\mathbf{s}$) indicate how much degree that modules in N_1 can be affected by condition D. The higher $\mathbf{s}_i$ means the module i in N_1 may just be responsible for general stress. Especially when some $\mathbf{s}_i$ in N_1 keeps relatively high row sum of A compared with all other N_2 (remove one condition each time), showing these modules have little association with any specific conditions. While lower $\mathbf{s}_i$ means module i in N_1 is very different from the modules in N_2, which may indicate the module has some connection with condition D. The rationale behind this simple criteria is based on the mechanism of correlation, i.e. which samples can make impact on the correlation coefficient while others may not? More details can be found in supplementary file part 1.

After determine which module may be condition specific, we can associate biological process with module by functional annotation enrichment analysis. The input can be gene list from the module, or overlapping just part much with others. Here we use DAVID [9] to conduct integrative functional annotation enrichment analysis of gene list based on an R Webservice interface [10]. We implemented a module differential analysis pipeline, from gene expression profile of multiple conditions to enrichment analysis results. Figure shows the general process of each step mentioned

above.

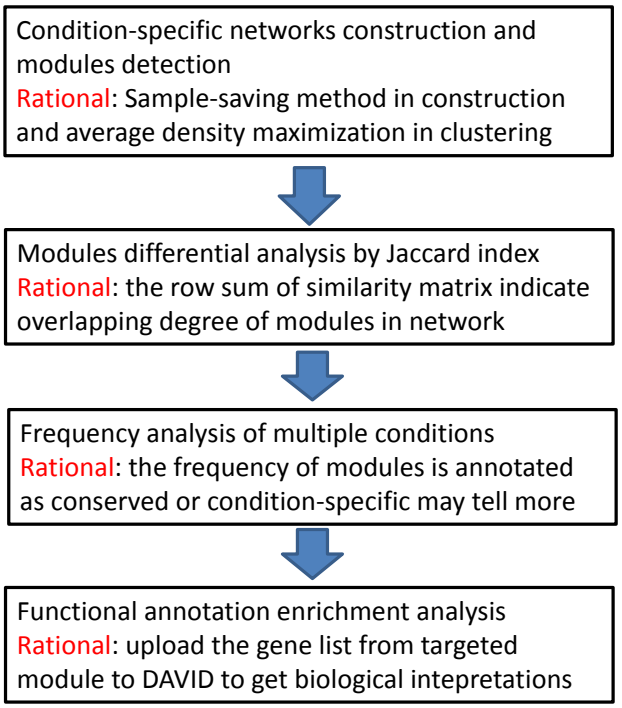

Figure 1: Overview of MODA.

4 Result

We evaluated the effectiveness of proposed methods on both synthetic data and real-world data. By comparing two gene expression profiles generated by different desired correlation matrices of the same set of genes, we can determine the genes affected by a groups definition, which is consistent with the generator. The details for simulation as well as the usage of package can be found in supplementary file part 2. The method is also used on a comprehensive RNA-Seq data set obtained from

two natural genotypes fo D. magna, to detect condition-specific as well as conserved responsive genes and biological functions. Several biological meaningful results show the capability of the algorithm, and more details can be found in [stressflea draft].

5 Supplementary

5.1 Concept part

Given gene expression profile $X \in \mathbb{R}^{n \times p}$, where n is the number of experimental samples and p is the number of genes. X_{ij} means the expression value of the j-th gene in i-th sample. The popular tool WGCNA [1] conducts the module detection by hierarchical clustering, i.e. putting similar gene together. The definition of similarity ranges from basic correlation to more complex topological overlap measure [2]. While how to determine the cutting height of hierarchical clustering tree remains an open problem. Here we give the option to chose the height based on the quality of partition. Inspired by the concept of partition density of link communities [3, 11], we choose the cutting height to make the average density of resulting modules to be optimal. The density of one module A is defined as:

$$Density(A) = \frac{\sum_{i \in A} \sum_{j \in A, j \neq i} a_{ij}}{n_A(n_A - 1)} \tag{1}$$

where a_{ij} is the similarity between gene i and gene j, and n_A is the number of genes in A. We can also use the modularity Q of weighted network A [4] as the criterion to pick the height of hierarchical clustering tree:

$$Q = \frac{1}{2m} \sum_{ij} [a_{ij} - \frac{k_i k_j}{2m}] \sigma(c_i, c_j) \tag{2}$$

where m is the number of edges and k_i is the connectivity (degree) of gene i, defined as $\sum_j a_{ij}$. And $\sigma(c_i, c_j) = 1$ only when gene i and j are in the same module. The complete module detection and average density is shown in Figure 2.

After the module detection, the co-expression network is represented as a collection of modules (see Figure 3), which makes the differential analysis more focused on the modules other than the nodes or links. By comparing all module pairs from N_1 and N_2, we can get a similarity matrix B, where each entry B_{ij} means the similarity between the i-th module from the network N_1 (denoted by $N_1(A_i)$) and j-th module from the network N_2 (denoted by $N_2(A_j)$). The similarity is evaluated by

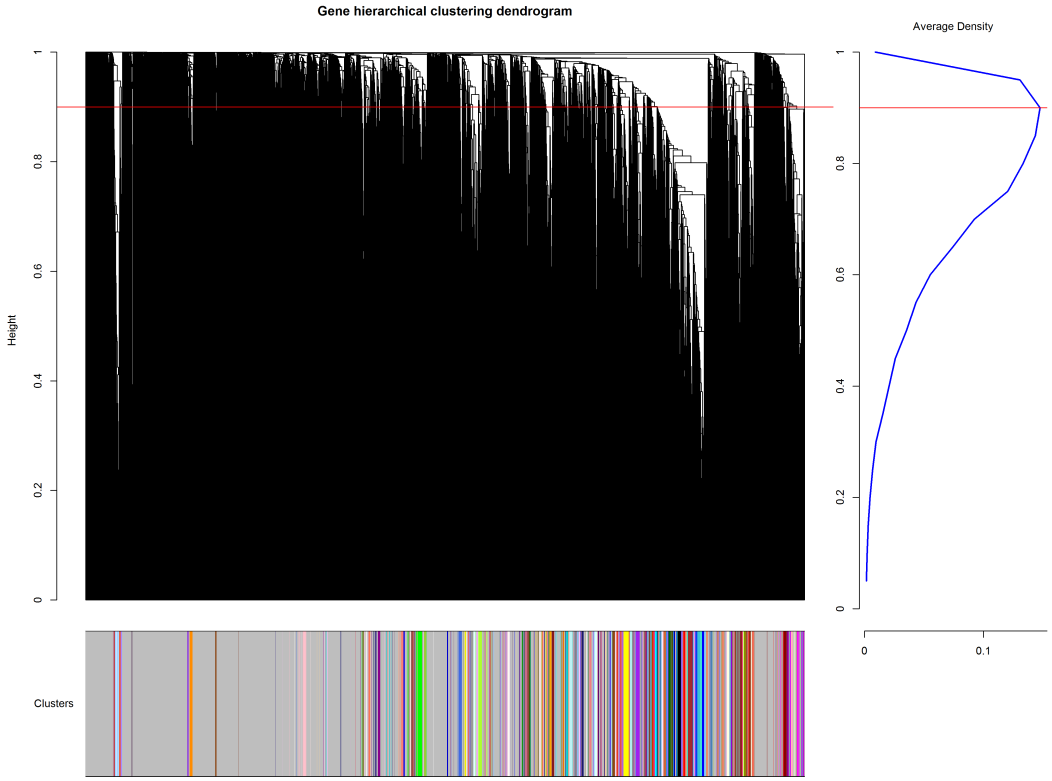

Figure 2: Maximal partition density based hierarchical clustering

the Jaccard index.

$$B_{ij} = \frac{N_1(A_i) \cap N_2(A_j)}{N_1(A_i) \cup N_2(A_j)} \tag{3}$$

Assume N_1 is background, normally containing samples from all conditions, and the N_2 is constructed from all samples except samples belonging to certain condition D. Let $\mathbf{s}$ is the sums of rows in B, i.e. $\mathbf{s}_i = \sum_j B_{ij}$. The value of $\mathbf{s}_i$ indicates how much the i-th module from network N_1 might be affected by condition D. The rationale behind this statistics is based on the mechanism of correlation, i.e. which samples could make an impact on the correlation while others may not? Figure 3 illustrates an extreme example about how the additional two samples may affect the correlation between X and Y.

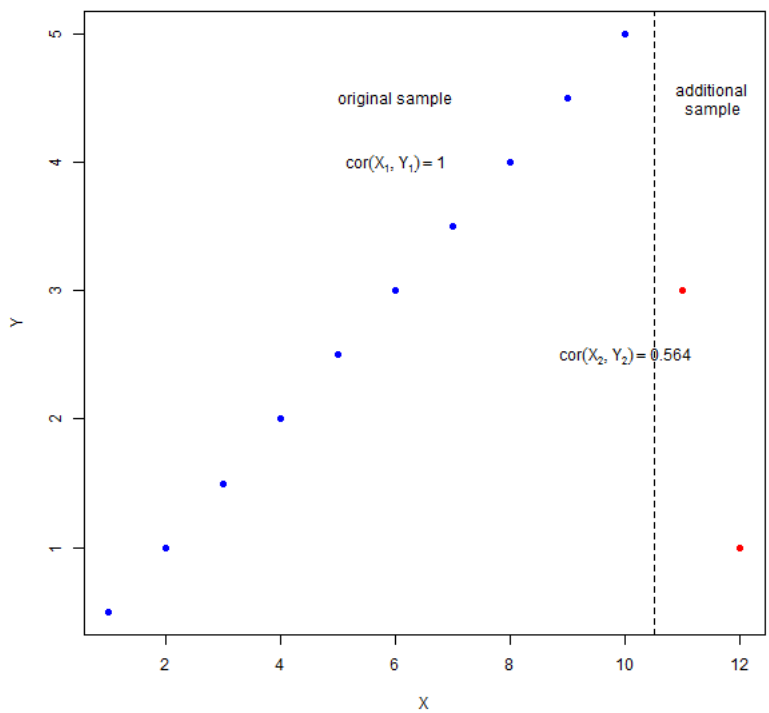

Figure 3: Scatter plot of varibale X and Y

As Figure 4 shows, we use two threshold values here: θ_1 is the threshold to define $min(\mathbf{s}) + \theta_1$, less than which is considered as condition specific module. θ_2 is

the threshold to define $max(\mathbf{s}) - \theta_2$, greater than which is considered as condition conserved module.

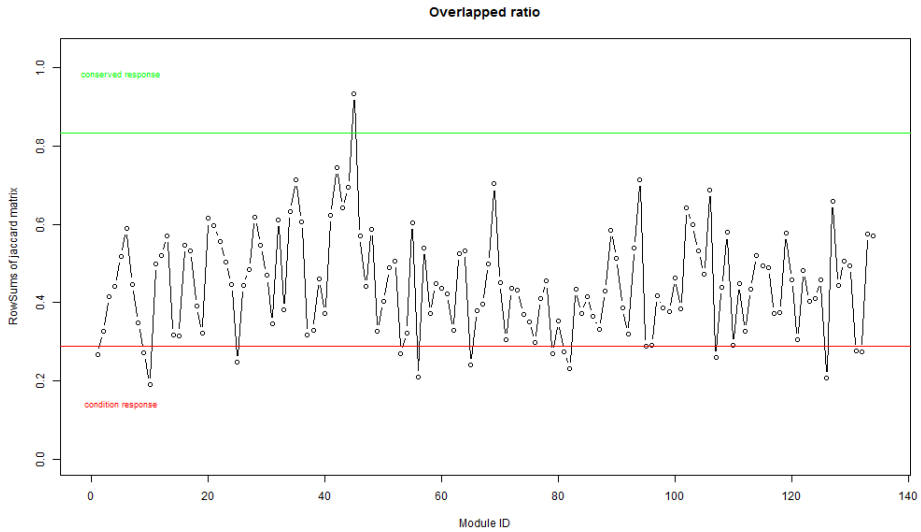

Figure 4: Overlap degree of modules in N_1 with N_2

We also calculate the frequency of each module is annotated as conserved or condition specific and compare all the conditions together. The rationale behind this statistics is based on the mechanism of correlation, i.e. which samples could make an impact on the correlation while others may not? The package visualizes it with a bar plot as Figure 4. A similar plot about the conserved module is also available. The module id is stored as a plain text file for functional enrichment analysis. Here we send one module as gene list to DAVID [9, 10] for integrative analysis.

5.2 Evaluation

We evaluate the effectiveness of proposed methods on both synthetic data and real-world data. The basic synthetic gene expression data is generated by the following logic: given desired correlation matrix $C \in \mathbb{R}^{n \times p}$ with p genes which has a clear modular structure that all genes are equally divided into 5 groups according to the similarities. Then we conduct the Cholesky decomposition on C such that $C = LL^T$, where L is the lower triangular matrix. Finally we project L on random matrix $A \in \mathbb{R}^{n \times p}$ to get desired gene expression matrix $X \in \mathbb{R}^{n \times p}$, which has the rough

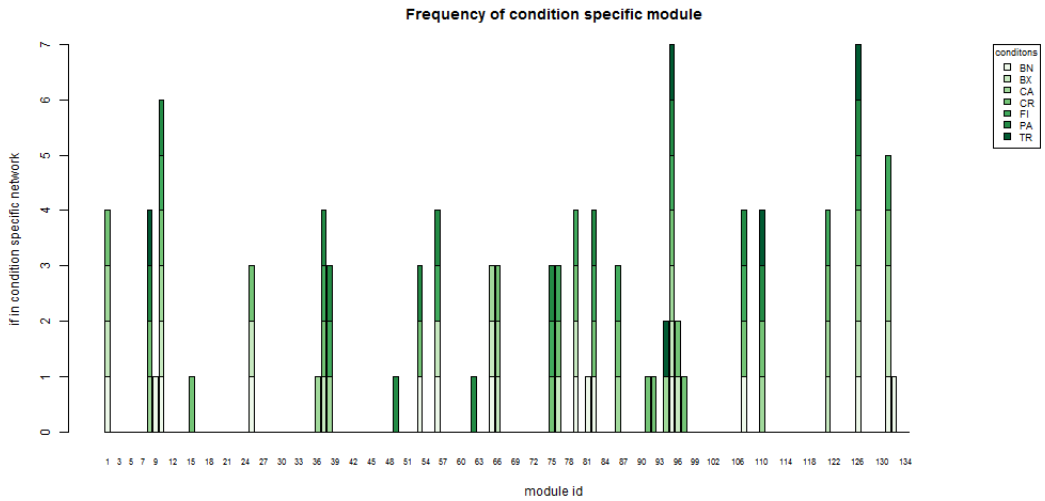

Figure 5: Statistics about which module can be condition specific

modular structure defined by correlation C. Let $n = 500$ and each group has 100 genes in the simulation. In each group, we allocate the gene id from 1-100, 101-200, 201-300, 301-400 and 401-500 respectively. The correlation matrix of genes in X is shown in Figure 6. In another matrix Y, we merge the last two groups into one by adding more samples to X, and the correlation matrix is shown in Figure 7. The we can compare these two networks with proposed method to see which genes were affected. Gene lists in target fold show that modules that contain gene id from 1-100, 101-200 and 201-300 have large overlap with network 2, while module gene id from 301-500 which were merged have least overlap with network 2. The facts are consistent with experimental settings.

Here is the example code to use `WeGonda` given two gene expression profiles. Results of modules are stored under the newly created folder *ResultFolder* as gene lists. The condition-specific and conserved module ids are stored as plain texts in next directory with the name of indicator which need to be compared. Other materials such as figure 2 and 4 are also available in the folder.

```
library('MODA')
ResultFolder = 'ForSynthetic' # where middle files are stored
CuttingCriterion = 'Density' #CuttingCriterion could be Density or Modularity
speciesName1 = 'X'       # indicator for data profile 1
speciesName2 = 'Y'       # indicator for data profile 2
specificTheta = 0.1 #threshold to define condition specific modules
conservedTheta = 0.1#threshold to define conserved modules

# modules for network 1
intModules1 ← WeightedModulePartitionDensity(datExpr1,ResultFolder,speciesName1,CuttingCriterion)
# modules for network 2
intModules2 ← WeightedModulePartitionDensity(datExpr2,ResultFolder,speciesName2,CuttingCriterion)
```

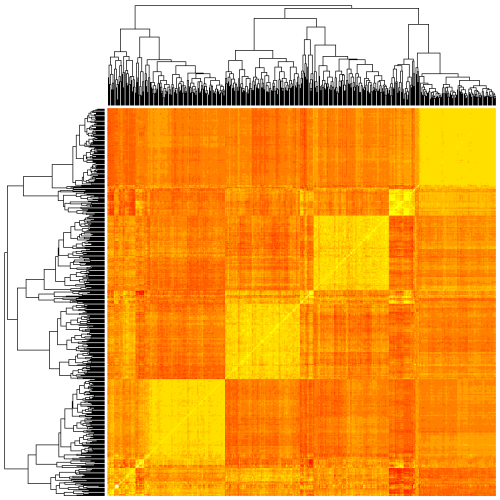

Figure 6: Correlation matrix of X

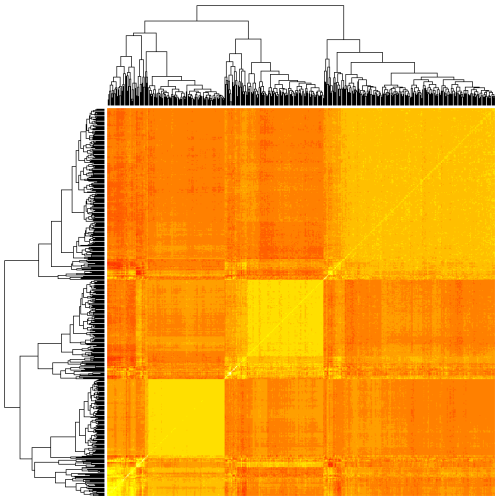

Figure 7: Correlation matrix of Y

```
# compare these two networks
CompareAllNets(ResultFolder,intModules,speciesName1,intModules2,speciesName2,specificTheta,conservedTheta)
```

References

[1] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[2] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[3] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[4] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[5] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[6] David Amar, Hershel Safer, and Ron Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol*, 9(3):e1002955, 2013.

[7] Gennaro Gambardella, Maria Nicoletta Moretti, Rossella de Cegli, Luca Cardone, Adriano Peron, and Diego di Bernardo. Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14):1776–1785, 2013.

[8] Marieke Lydia Kuijjer, Matthew Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *arXiv preprint arXiv:1505.06440*, 2015.

[9] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2008.

[10] Cristóbal Fresno and Elmer A Fernández. Rdavidwebservice: a versatile r interface to david. *Bioinformatics*, page btt487, 2013.

[11] Alex T Kalinka and Pavel Tomancak. linkcomm: an r package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*, 27(14), 2011.