ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 18(4): 138-441 (2022)

©Biomedical Informatics (2022)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 18(4)

Research Article

Received March 1, 2022; Revised April 30, 2022; Accepted April 30, 2022, Published April 30, 2022

DOI: 10.6026/97320630018438

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Edited by P Kangueane
Citation: Kumar *et al.* Bioinformation 18(4): 438-441 (2022)

Linking co-expression modules with phenotypes

Rakesh Kumar[1], Niharika[1], Krishna Kumar Ojha[1], Harlokesh Narayan Yadav[2] & Vijay Kumar Singh[1]*

[1]Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar 824236, India; [2]Department of Pharmacology, All India Institute of Medical Sciences, Ansari Nagar, New Delhi - 110029, India. *Corresponding author: Vijay Kumar Singh - Email: vksingh@cub.ac.in

Abstract:
The method for quantifying the association between co-expression module and clinical trait of interest requires application of dimensionality reduction to summaries modules as one dimensional (1D) vector. However, these methods are often linked with information loss. The amount of information lost depends upon the percentage of variance captured by the reduced 1D vector. Therefore, it is of interest to describe a method using analysis of rank (AOR) to assess the association between module and clinical trait of interest. This method works with clinical traits represented as binary class labels and can be adopted for clinical traits measured in continuous scale by dividing samples in two groups around median value. Application of the AOR method on test data for muscle gene expression profiles identifies modules significantly associated with diabetes status.

Keywords: Analysis of rank, co-expression, gene module, gene expression, network.

Background:

In recent years the transcriptomic data analysis has witness a shift from gene level analysis to gene modules level analysis **[1, 2]**. Module level analysis aims to identify set of co-expressed/co-regulated genes from transcriptomic data **[3, 4]**. Further downstream analysis includes identification of intra-modular hubs, investigation of relationship between co-expression modules and the comparison of network topology of different networks **[5, 6]**. Generally module as whole is summarized with either meta-gene representing average expression of genes for each sample or with module eigen genes which can be considered as the best summary of standardized module expression data **[7]**. The module eigen gene of a given module is defined as the first principal component of the standardized expression profiles **[8]**. To find modules that relate to a clinical trait of interest, the module eigen genes are correlated with the clinical trait of interest **[9, 10]**. Therefore, it is of interest to describe a method using analysis of rank (AOR) to assess the association between module and clinical trait of interest.

Materials and methods:

The AOR algorithm to assess module-trait association:

The method works such that where clinical traits are distinct binary class (positive or negative). Given the total number of samples is m and out of which k belongs to negative class, the method examines the expression pattern of modules across the samples and identify modules which tend to have significant association with the clinical trait of interest. The method is based on the observation that the samples of the two class will show clear distinction in distribution of ranks if they are arranged according to expression values of a gene having true differential expression. However, in module level analysis instead of a single gene we deal with set of genes assigned to the module. Therefore, in order to find sample ranking based on expression of modules as whole, a matrix with n rows, equal to the number of genes in the module and m columns, equal to the number of total samples was created. Each row of the matrix contains samples arranged in ascending order as per the expression values of genes. A position vector was created by calculating the negative class sample frequency for each column of the matrix. For a module which is not related to clinical trait, the first k largest frequencies will be uniformly distributed across the position vector. However, a module having significant relation with clinical trait will cause larger frequencies to concentrate towards one end of the position vector. The index of first k largest frequencies were summed to get a score G_s. A significantly lower G_s score represents lower expression of module in negative class sample and a significantly higher score represents higher expression of module in negative class sample.

Calculation of score G_s

A module M was defined as collection of genes $g_i(i=1...n)$ having similar expression pattern across the samples $s_j(j=1...m)$. The expression information of $g_i \in M$ across samples can be stored in a nxmmatrix E where value E_{ij} represents the expression information of i^{th} gene in j^{th} sample. A sample s_j was assigned to set L^0 if it belongs to negative class and to set L^1 if belongs to positive class. The expression matrix E was converted to nxmindex matrix I where value I_{ij} was set as per Eq. (1).

$$I_{ij} = \begin{Bmatrix} 0 if O_j^s \in L^0 \\ 1 if O_j^s \in L^1 \end{Bmatrix} \qquad (1)$$

Where O^scontains samples arranged in ascending order according to expression values of g_i. Each columns of index matrix was summed to create a position vector v

Where, $v_j = \sum_{i=1}^{n} I_{ij}$

The score G_s was calculated by adding indexes of the first k largest values in the position vector v.

Assessing the significance of score G_s

A null distribution of score G_s was calculated by assuming that the module is not associated with the clinical trait. In case of no association, the position vector v will have uniform distribution of first k largest values. Therefore, in order to assess the significance of G_s score of a module having n number of genes following steps were performed:

[1] False modules were created by randomly selecting n genes from the total number of genes for which expression data is available.

[2] The score G_s for False module was calculated.

[3] Step 1 and 2 were repeated 1000 time to generate a null distribution of score G_s.

A module with G_s score greater than $|\mu \mp (1.96xS)|$ was considered significantly associated with negative class samples. Where μ and S are mean and standard deviation of the null distribution respectively. The score $G_s$was further converted to scaled $G_s$score (G_S^{scaled}) and he standardised $G_s$score ($G_S^{stan\,derdised}$). The score $G_s$was scaled to have value between -1 and +1 using Eq. (2).

$$G_S^{scaled} = 2 * \frac{G_S - G_S^{min}}{G_S^{max} - G_S^{min}} - 1 \qquad (2)$$

Where G_S^{min} and G_S^{max} are the minimum and maximum permissible value for score G$_s$. The $G_S^{standerdised}$ was calculated by subtracting the oserved G_s score from the mean of the null distribution and dividing by the standard deviation of the null distribution.

Results:

The R-package WGCNA **[11]** was used to identify set of co-expressed genes from muscle transcriptome of healthy (NGT) and diabetes (T2D) subjects **[12]**. In total WGCNA has identified 30 modules having tightly co-expressed genes grouped together. In order to assess the association between modules and subject diabetes status we applied the AOR and module Eigen gene (as implemented in WGCNA package) method to expression data of each of the module. Using the cutoff of p value < 0.001 thirteen modules were found significantly associated with subject diabetes status using AOR method whereas module Eigen gene method was not able to produce statistical significant result for any of the module (Figure 1). This showed that the subject diabetes status have significant affect on muscle gene expression.

Discussion:

In this study, analysis of rank (AOR) method has been used to assess the association between clinical traits and co-expression modules. Application of the method on muscle gene expression profile identified significant association between module and subject diabetes status highlighting importance of AOR method in identifying hidden patterns across gene expression profiles. While interpreting the results of the AOR method both type of score G_S^{scaled} and $G_S^{standerdised}$ should be taken into consideration along with the obtained p-value. The significant p-value obtained using AOR method for module-trait association just indicates that, more number of negative class samples, than expected by chance, are concentrated at the beginning/end of the position vector. The similar sign for $G_S^{standerdised}$ and G_S^{scaled} score for a module supports deregulation of genes belonging to concerned module between negative and positive class samples. However, opposite signs for $G_S^{standerdised}$ and G_S^{scaled} suggest the existence of expression heterogeneity within negative class samples with respect to expression of genes belonging to concern module.

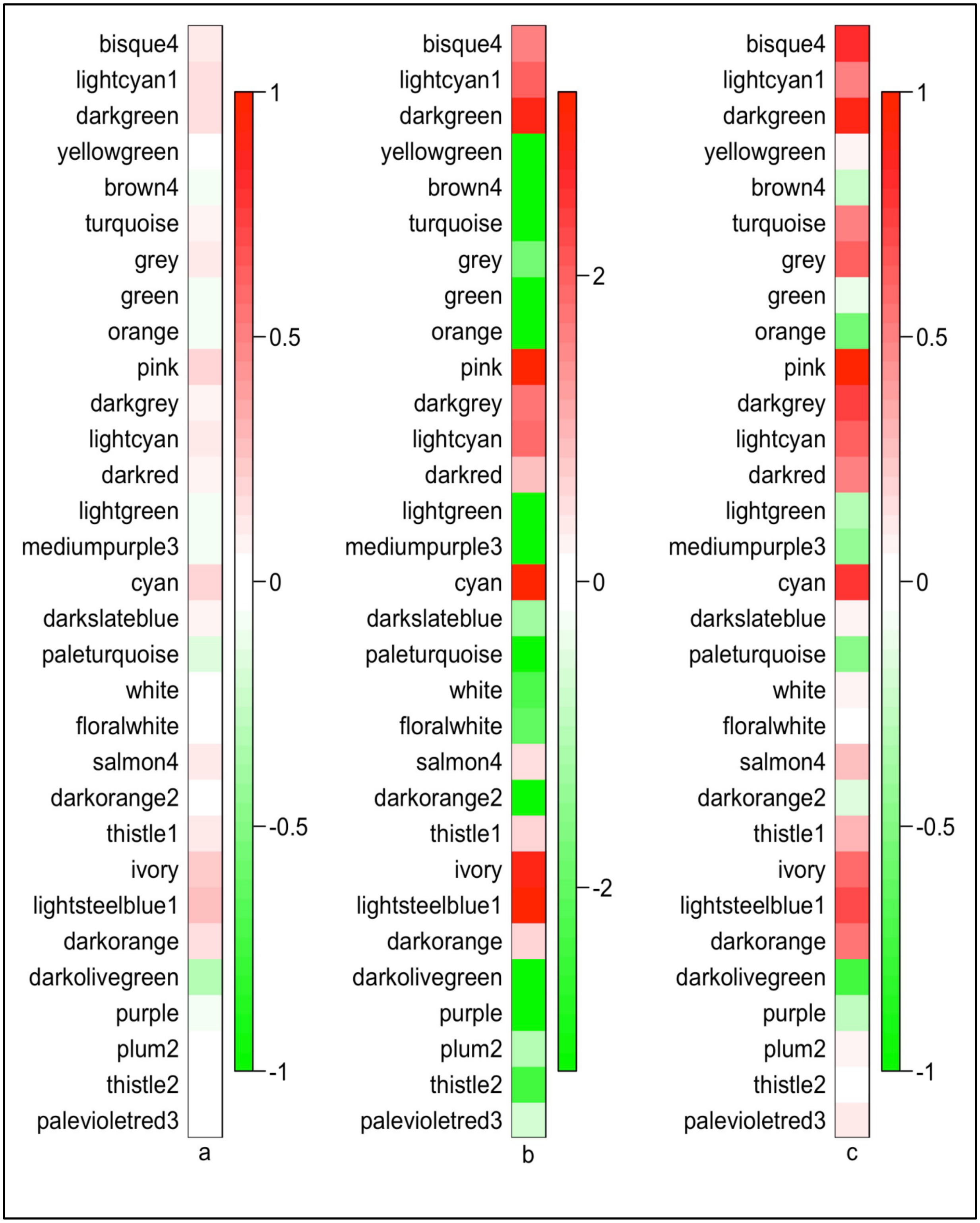

Figure 1: Heatmap representation of module-trait correlation matrix. The modules were identified from muscle expression data of normal and diabetic subjects using WGCNA R-package. The colour represents the extent of correlation between module and subject diabetes status (clinical traits). a) The module-trait association as quantified using eigengene method as available in WGCNA package. b) The module-trait association as quantified using AOR method. Cells were coloured as per standardised Gs score and c) The module-trait association as quantified using AOR method. Cells were coloured as per scaled Gs score.

Conclusion:

The AOR method helps in identifying hidden patterns from gene expression data and can provide deeper insights into disease biology by discovering co-expression modules linked to clinical-traits.

Conflict of interest:

The author(s) declared no conflicts of interest with respect to the research, authorship, and publication.

Acknowledgements:

We thank University Grant Commission (UGC), Ministry of Education, Government of India, for providing research fellowship (SRF) to Mr. Rakesh Kumar under Rajeev Gandhi National Fellowship (RGNF) scheme. This work was supported by the Indian Council of Medical Research (ICMR) extramural grant [Project number: 2020-3854] and Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India, young scientist grant [No: YSS/2015/01759].

References:

[1] Choobdar S *et al. Nature Methods* 2019 **16**:843 [PMID: 31471613]

[2] Saelens W *et al. Nature Communications* 2018 **9**:1 [PMID: 29545622]

[3] Mao Y *et al. Molecular Medicine Reports* 2020 **22**:1155 [PMID: 32468072]

[4] Wang J & Xia S, *PLoS Computational Biology* 2016 **12**:e1004892 [PMID: 27100869]

[5] Chowdhury HA *et al. IEEE/ACM transactions on Computational Biology and Bioinformatics* 2019 **17**:1154 [PMID: 30668502]

[6] Van Dam S *et al. Briefings in Bioinformatics* 2018 **19**:575 [PMID: 28077403]

[7] Horvath S & Dong J, *PLoS Computational Biology* 2008 **4**:e1000117 [PMID: 18704157]

[8] Langfelder P & Horvath S, *BMC Systems Biology* 2007 **1**:1 [PMID: 18031580]

[9] Langfelder P & Horvath S, *BMC Bioinformatics* 2008 **9**:1 [PMID: 19114008]

[10] Song WM & Zhang B, *PLoS Computational Biology* 2015 **11**:e1004574 [PMID: 26618778]

[11] Langfelder P & Horvath S, *J Stat Softw* 2012 **46**:i11 [PMID: 23050260]

[12] Gallagher IJ *et al. Genome Medicine* 2010 **2**:1 [PMID: 20353613]

Agro Informatics Society
BIOMEDICAL INFORMATICS
BMIS
AGIS
Biomedical Informatics Society
since 2005
BIOINFORMATION
Discovery at the interface of physical and biological sciences
indexed in
PMC
PubMed
EBSCO
EMERGING SOURCES CITATION
WEB OF SCIENCE
Web of Science Group
Crossref
doi
ResearchGate
RG
publons