Hittite Journal of Science and Engineering, 2019, 6 (4) 315-318
ISSN NUMBER: 2148-4171
DOI: 10.17350/HJSE19030000163

A modified Soft-thresholding Approach in the Transcriptomic Analysis of Adaptation of E.coli to Alternating Substrate Conditions

Muhammed Erkan Karabekmez
İstanbul Medeniyet University,Department of Bioengineering, İstanbul, Turkey

ABSTRACT

The expression of genes that are functionally related is considered to change together in response to deterioration of internal or external order. The system-level analysis of these changes has become widespread in recent years. Weighted gene co-expression network analysis (WGCNA) is an important tool in the literature. This method has two options in the form of hard and soft thresholding. The power function is used commonly in soft thresholding option. The other alternative of soft thresholding, symmetric sigmoid function, may give less importance to the meaningful co-expression data and not preferred frequently. Both functions has some drawbacks. In this study, it was tried to increase the efficiency of WGCNA approach by using asymmetric sigmoid function. RNA-seq dataset on adaptation of E.coli to alternating substrate conditions was re-investigated with this modified approach and its use was proven by GO and pathway enrichment analysis.

Keywords:
WGCNA, Transcriptomics, Asymmetric Sigmoid Function

Article History:
Received: 2019/11/07
Accepted: 2019/11/20
Online: 2019/12/31

***Correspondence to**: Muhammed Erkan Karabekmez*
İstanbul Medeniyet University,
Department of Bioengineering
Tel: o(216) 280 3333
E-Mail:
erkan.karabekmez@medeniyet.edu.tr

INTRODUCTION

Computational techniques have been developed to reveal the biological significance of the large transcriptomic datasets with the emergence of microarray technique in 2000s, which allows high-throughput measurement of gene expression data at the genome level. These techniques are largely based on clustering algorithms, mathematical modeling and network analysis. In addition to the applications of a wider range of mathematical approaches, there are also mathematical approaches developed specifically for this biological context. The most common of these approaches is Zhang and Horvath's weighted gene co-expression analysis (WGCNA) approach, published in 2005 [1]. This method was later improved by many studies [2-9], expanded for meta-analysis [10] and R ready-to-run software was also provided [11]. The WGCNA approach has been used in many studies and has received thousands of citations [12-16].

First, WGCNA numerically calculates the correlations of gene expression across different conditions or temporal points between each gene pair. It then thresholds these pairwise correlation coefficients to transform them into discrete values - which is defined as solid thresholding - or uses the force function or the sigmoid function to inflate high attenuation and weaken weak attenuation, which is called soft thresholding [1].

In the WGCNA approach, α (when sigmoid function is used), β (when force function is used), and τ (when hard-thresholding used) variables are determined in a way that ensure the network will be independent of the scale, because biological networks are known to be scale-free [17].

By using sigmoid function for transformation, gene pairs whose correlation coefficient is slightly higher than the saddle point, which supposed to be 0.5 are overrated that is why power function is much more common in the literature. On the other hand, by using power function, correlation coefficients that are slightly lower than 1.0 are underrated. Intuitively, it can be hypothesized that shifting saddle point of sigmoid function upwards, i.e. asymmetric sigmoid function, can be used for transformation to avoid both of the drawbacks. However, asymmetric sigmoidal functions have more than one coefficients which complicates the parameterization step. Here in this study we attempted to develop a pipeline to use asymmetric sigmoidal function for soft-thresholding in WGCNA. We used an RNAseq

This paper was presented orally in 2nd International Conference on Life and Engineering Sciences, ICOLES 2019
27th - 29th of June 2019, Medipol University, Istanbul, Turkey

dataset on adaptation of E.coli to alternating substrate conditions to validate this novel approach.

MATERIALS AND METHODS

RNA-seq dataset of E.coli was retrieved from GEO with the accession code of GSE97944 [18]. Quantifications of the dataset was in terms of fragments per kilobase per million (FPKM) and obtained by cufflinks [19] following alignment by using Bowtie 2 algorithm [20].

Cytoscape version 3.6.1 was used to visualize networks [21].

MATLAB R2018b platform was used for thresholding calculations.

When clustering with the WGCNA approach, the existing package on the R platform was used [11].

By using gene ontology and pathway enrichment analyzes, the biological significance of the resulting modules was examined. DAVID 6.8 web-based software tool was used for this purpose [22].

RESULTS

RNA-seq dataset of E.coli is composed of expression levels of 3754 genes at 8 conditions. One of the advantages of WGCNA approach is making use of the whole dataset instead of reducing to a subset of differentially expressed genes. First of all for each gene pair a distance metric (Pearson Correlation Coefficient) is calculated to quantify the similarities between expression profiles accross 8 conditions. Secondly, a thresholding approach is followed to reduce the noise. Three basic thresholding approaches- hard thresholding, soft thresholding with power function and soft thresholding with sigmoid function- were used to assess the scale-free networks. Transformation functions with different parameters of sigmoid function (Fig. 1A) and power function (Fig. 1B) were plotted. Sigmoid function provides little transformation while power function leads a huge difference between high similarities.

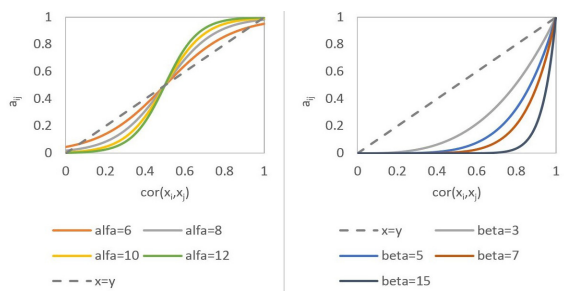

Figure 1. Sigmoid functions with various α values (A) and power function with various β values (B)

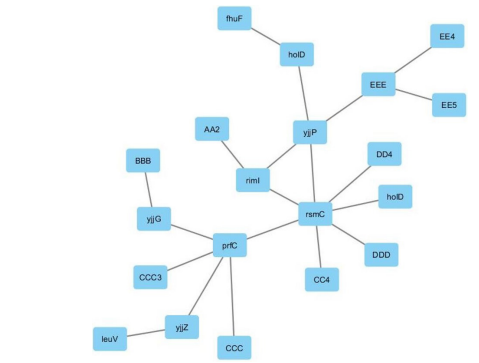

Figure 2. A typical scale-free network with a degree-distribution that fit to a linear decrease of frequency with increasing degree on log-scale (R2 = 0.918).

In scale-free networks degree distribution follows power law distribution that is why it should be linear in logarithmic scale. If a line fit to a distribution by linear regression with a slope less than -0.5 and an R2 value higher than 0.7 the degree distribution was accepted as scale-free. Hereby one can conclude that the constructed network is biologically relevant. A typical scale-free network were plotted by using Cytoscape in order to display the structure (Fig. 2).

It was found that minimum τ value for hard-thresholding was 0.76 to attain scale-free criteria and minimum β value for soft-thresholding with power function was 7. Scale-free criteria could not be attained with sigmoidal function with any parameter.

As a forth thresholding strategy, here in this study, asymmetric sigmoidal function with two parameters (Eq. 1) was used as a novel transformation function.

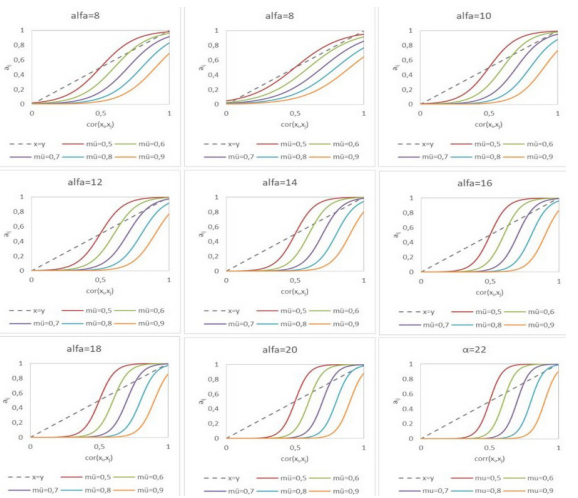

Figure 3. Effect of parameterization on transformation by asymmetric sigmoid function

$$a_{ij} = \frac{1}{1 + e^{-a(|cor(x_i x_j)| - \mu)}} \quad (1)$$

The effect of different values of α and μ were simulated to display the transformation that they can cause (Fig. 3).

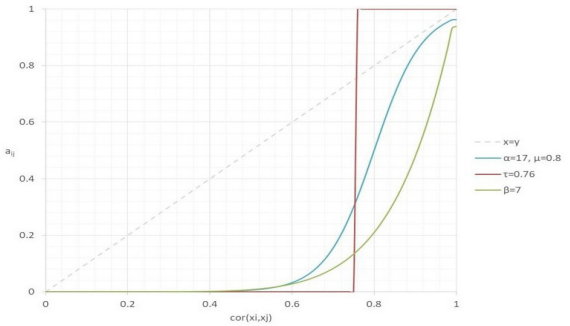

Figure 4. Transformations by hard-thresholding (step-function) (red curve) and soft thresholding by using conventional power function (green curve) and novel asymmetric sigmoidal function (blue curve)

Two hundred different combinations of α and μ were searched by grid search approach and scale-free nature of resulting network for each combination was investigated. The lowest possible μ value was picked first and lowest possible α in combination with the fixed μ was chosen. As a result 0.8 for μ and 17 for α parameters were identified to be attaining scale-free criteria (Fig. 4).

Resulting modules calculated by using the parameterized asymmetric sigmoidal function through the R-package. Biological significance of the modules identified by using power function and asymmetric sigmoidal function were compared with respect to gene ontology and pathway enrichments.

It was observed that larger modules with finer association to specific biological roles were attained by using asymmetric sigmoidal function. For instance; KEGG pathway ribosome was found to be associated with a module with a size of 40 genes with a p-value of 2.42E-04 with power function whereas there is a larger module of 56 genes with a more significant association to the same pathway with a p-value of 1.65E-13.

Resulting enrichment analysis were also showed that genes involved in following biological processes and pathways mediates adaptive response of E.coli to alternating substrate conditions; oxidative phosphorylation, biosynthesis of amino acids, flagellar assembly and response to stress.

DISCUSSION

WGCNA is one of the most common approaches for transcriptomic data analysis. It is shown in this study that asymmetric sigmoidal function with two parameters can improve performance of WGCNA and its usefullness was shown in adaptive response of E.coli to alternating substrate conditions. The parameterization process should be validated across different datasets and a systematic procedure should be developed to standardize the future effords. And as another further work this approach should be included into automated R-packages of WGCNA for its wider use.

ACKNOWLEDGEMENT

This work was supported by Istanbul Medeniyet University through BAP Grant No. F-GAP-2018-1245.

REFERENCES

1. Zhang B and Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, (2005) Article 17 PMID: 16646834.
2. Dong J, Horvath S. Understanding Network Concepts in Modules. BMC Systems Biology, 1:24 (2007) PMID: 17547772 PMCID: PMC3238286.
3. Horvath S, Dong J. Geometric Interpretation of Gene Coexpression Network Analysis. PLoS Comput Biol 4(8): (2008) e1000117 PMID: 18704157 PMCID: PMC2446438.
4. Yip A, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics, 8:22. (2007) PMID: 17250769 PMCID: PMC1797055.
5. Li A, Horvath S. Network Neighborhood Analysis with the multi-node topological overlap measure. Bioinformatics. (2006) PMID: 17110366 doi:10.1093/bioinformatics/btl581.
6. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R. Bioinformatics. November/btm563 (2007) PMID: 18024473.
7. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Systems Biology 2008, 2:34 (2008) PMID: 18412962 PMCID: PMC2387136.
8. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PloS Comp Biol. 7(1): (2011) e1001057 PMID: 21283776 PMCID: PMC3024255.
9. Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinformatics 14:5 (2013) PMID: 23323760 DOI: 10.1186/1471-2105-14-5.
10. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54 (2007) PMID: 18031580.
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9:559. (2008) PMID: 19114008 PMCID: PMC2631488.
12. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu JY, Horvath S, Fan G. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013 Jul 28. (2013) doi: 10.1038/nature12364 PMID: 23892778.
13. van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, de Kovel CG, Janson E, Strengman E, Langfelder P, Kahn RS, van den Berg LH, Horvath S, Ophoff RA. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects BMC Genomics. 2012 Nov 17;13(1):636 (2012) PMID: 23157493.
14. Wang, W., Jiang, W., Hou, L., Duan, H., Wu, Y., Xu, C., ... & Zhang, D. Weighted gene co-expression network analysis of expression data of monozygotic twins identifies specific modules and hub genes related to BMI. BMC genomics, 18(1), 872. (2017).
15. Miller, J. A., Guillozet-Bongaarts, A., Gibbons, L. E., Postupna, N., Renz, A., Beller, A. E., ... & Szafer, A. Neuropathological and transcriptomic characteristics of the aged brain. eLife, 6. (2017).
16. Simon, S., Sagasser, S., Saccenti, E., Brugler, M. R., Schranz, M.

E., Hadrys, H., ... & DeSalle, R. Comparative transcriptomics reveal developmental turning points during embryogenesis of a hemimetabolous insect, the damselfly Ischnura elegans. Scientific Reports, 7(1), (2017) 13547.

17. Albert, R., & Barabási, A. L. Statistical mechanics of complex networks. Reviews of modern physics, 74(1), (2002), 47.
18. Sandberg, T. E., Lloyd, C. J., Palsson, B. O., & Feist, A. M. Laboratory evolution to alternating substrate environments yields distinct phenotypic and genetic adaptive strategies. Appl. Environ. Microbiol., 83(13), (2017), e00410-17.
19. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L.. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5), 511, (2010).
20. Langmead, B., & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357, (2012).
21. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504, (2003).
22. Huang, D. W., Sherman, B. T., & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), (2008), 44.