Development of an Automated Biological Tool for Visualizing Dissimilarity within Gene Co-Expression Networks in Hierarchical Clustering

Prissadang Suta*, Panissara Thanapol†, Jonathan H. Chan‡, and Thiptanawat Phongwattana§
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Email: *prissadang.sut@mail.kmutt.ac.th, †panissara.tha@mail.kmutt.ac.th, ‡jonathan@sit.kmutt.ac.th,
§thiptanawat.p@mail.kmutt.ac.th

***Abstract*—We present an automated biological tool in order to visualize gene co-expression from a gene expression dataset in form of a dendrogram of hierarchical clusters. Our proposed tool calculates dissimilarities within gene-sets of biological pathways using Topological Overlap Measure algorithm. There are two aspects in our motivation comprising an automated biological tool and algorithms that can be utilized for researchers who are interested in this area because it can help to save time in the data preparation stage. Currently, WGCNA that is developed in R language is still limited in some processes that a researcher has to do manually, for instance data pre-processing and visualization. However, the library can be utilized in gene pair correlation computation of a gene co-expression that is calculated between each gene pair in a gene-set. Moreover, the popular library is able to transform the networks into scale-free networks in order to calculate the dissimilarity weights that are used to create a gene-set profile. In this work, we combined all of the necessary steps in form of an automated tool. Furthermore, our approach also distinguishes between gene pairs consisting of one, both, or no statistical significant genes, based on ANOVA testing of a set of features. In conclusion, our automated tool provides a means of clustering visualization in terms of biological pathway, as well as how gene dissimilarity is linked to the mutual significance of gene pairs within a gene-set that can help researchers in relevant fields to analyze their data.**

I. Introduction

At present, there are many software libraries implemented in the biological domain, especially in form of molecular data such as genes. A popular library for gene network analysis that we found is called "WGCNA" [1], which stands for weighted gene co-expression network analysis, developed in R language. The library can be utilized to find gene co-expressions by using Pearson correlation algorithm. Furthermore, it is able to create a scale-free network between genes. At the end of WGCNA process, it outputs a set of dissimilarities of each gene pair with in a gene-set that is part of a biological pathway, which refers to a disease. According to the prior research, we found that the outputs can be plotted into a graph in order to observe its characteristic that may lead to identification of potential biomarkers. The WGCNA library provides the core process but there is still a need to manually perform tasks such as data pre-processing and visualization. In the data pre-processing, source data, which is obtained from gene expression omnibus (GEO), which is a database repository of high throughput gene expression data and hybridization arrays, chips, and microarrays, is in a form of a probe dataset that is necessary to be annotated to generate a gene expression dataset. This may be a time-consuming step for researchers. Also, visualization is a necessity in biological field that the prior research needs to plot manually for each pathway of interest. According to the abovementioned issues, our proposed tool can help researchers to perform more efficient analysis as the tool just needs a raw probe dataset and the annotation file for its initialization. Then it can automatically execute the pipeline to produce visualization output in form of a hierarchical dendrogram. Moreover, we use a statistical technique called "ANOVA" in order to distinguish between gene pairs that consist of significant and non-significant label for each gene in a gene-set. By doing so, a set of the statistical outputs can be utilized as potential features for research. We also provide hyper-parameters for researchers in order to perform fine-tuning of the correlation algorithms, e.g. the number of power term for scale-free networks, complex disease pathways, and so on. For our contribution, researchers can leverage our developed tool for reducing their analysis time for to obtain analytic results since the tool has merged all necessary techniques and algorithms into a single pipeline. And the researchers can utilize our tool for parameter tuning in order to find some significances in the area of interest more easily.

II. Methodology

Lung cancer gene expression data (GSE10072) is downloaded from Gene Expression Omnibus (GEO) database [2]. GSE10072 dataset consists of 107 samples that can be categorized into 2 groups comprising Adenocarcinoma, which has 58 samples, and Non-tumor, which has 49 samples that are classified as control [3]. The samples were taken from tissue samples of Adenocarcinoma that was paired with non-involved lung tissue from current, former and non-smokers.

In Fig. 1, we provide the proposed method which is able to identify biomarkers by using gene co-expression datasets.

978-616-455-375-0

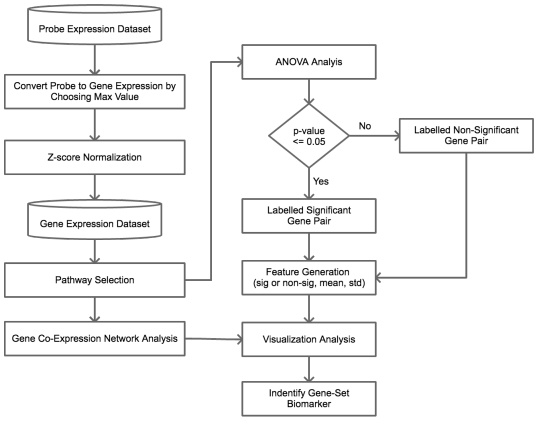

Fig. 1. Flowchart for identify potential biomarker.

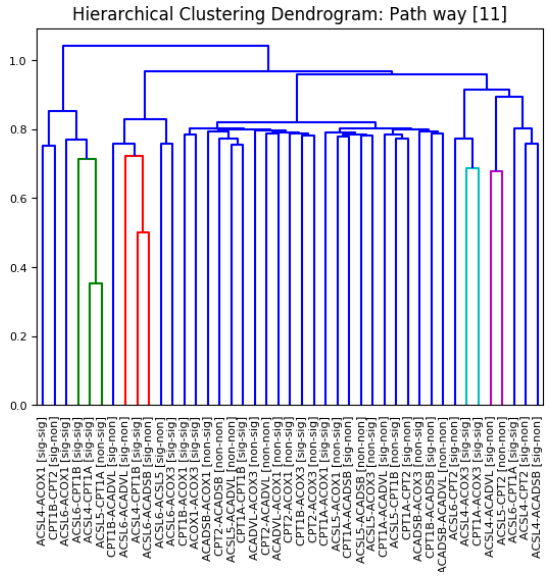

Fig. 2. Dendrogram of Hierarchical Clustering.

To identify each gene or biological pathway for gene-set transformation, it is divided into two groups by reference to ANOVA in order to tag labels significant (sig) or non-significant (non) in each gene-pair and categorize as sig-sig, sig-non and non-non, which is based on a threshold that is defined p-value $<= 0.05$. Next, we perform the gene co-expression network analysis. It includes Pearson's correlation, that generates a power term which the WGCNA is used as a default, i.e. beta equals to 6 that works well to analyze in gene co-expression networks [4]; it then creates a scale-free network topology [5], computes dissimilarity between genes using topological overlap measure dissimilarity [6], generates hierarchical clusters of genes, divides clustered genes into modules, and merges very similar modules. Finally, the measured dissimilarity will be used for graph plotting and it can be utilized for analyzing least errors that can be described how gene-pairs are visualized within gene co-expression networks. The visualization illustrates a graph characteristic and this result will be used to analyze by domain experts for biomarker identification. The result can identify biomarkers that are automatically obtained from gene co-expression visualization tool.

III. Results

In Fig. 2, we demonstrate a hierarchical dendrogram for the Fatty Acid Metabolism pathway. It consists of gene pairs that are categorized into each cluster. The value of each gene pair is based on weighted dissimilarity calculation that we use for clustering. However, some set of values in some pathway may have negligible variance, for example, they are very close to 0 or 1 mostly, and we would like to increase their spatial distribution in order to analyze clustering more effectively. The consequence is adding some exponential number to all of the values before categorizing into clusters. We also find a set of significant genes in each gene profile, which uses ANOVA, in order to analyze the significant genes (e.g. significant-significant, nonsignificant-significant, and nonsignificant-nonsignificant) in each cluster of dissimilarity of gene pairs for finding some potential relevance. Nonetheless, the statistical information in each cluster is still ambiguous. So, we would leverage the features in our future work for biomolecular analysis. Furthermore, we will colorize the clusters to highlight the hierarchical clustering visualization.

IV. Conclusion

This research paper contributes a useful biological tool that helps researchers to reduce lead time in research, especially in bio-data pre-processing. The automated tool can perform gene expression annotation, significance tagging, gene correlation analysis, weighted scale-free network analysis, topology overlap measure calculation, dissimilarity of gene co-expression network analysis as well as hierarchical clustering visualization within a single pipeline. In each stage, we also provide hyper-parameters tuning for convenience.

Acknowledgment

We would like to thank Narumol Dougpan, who gave us very useful advice and information for developing our automated tool that is based on WGCNA.

References

[1] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[2] T. Zeng, S. Y. Sun, Y. Wang, H. Zhu, and L. Chen, "Network biomarkers reveal dysfunctional gene regulations during disease progression," *FEBS Journal*, vol. 280, no. 22, pp. 5682–5695, 2013.

[3] T. Barrett, "NCBI GEO: mining millions of expression profiles–database and tools," *Nucleic Acids Research*, vol. 33, pp. D562–D566, 2004.

[4] M. T. Landi and et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, 2008.

[5] A. Barabási, E. Ravasz, and T. Vicsek, "Deterministic scale-free networks," *Physica A: Statistical Mechanics and its Applications*, vol. 299, no. 3–4.

[6] A. M. Yipand and S. Horvath, "The generalized topological overlap matrix for detecting modules in gene networks," pp. 1–19, 2005.