Gene-Set Profiles: Visualizing Dissimilarity Within Gene Co-Expression Networks for Biomarker Identification

Saila Shama
University of Toronto
Toronto, Canada
saila.shama@mail.utoronto.ca

Philip Lu
University of Toronto
Toronto, Canada
philip.lu@mail.utoronto.ca

Narumol Doungpan
KMUTT
Bangkok, Thailand
narumol.dou@sit.kmutt.ac.th

Asawin Meechai
KMUTT
Bangkok, Thailand
asawin.mee@kmutt.ac.th

Jonathan H. Chan
KMUTT
Bangkok, Thailand
jonathan@sit.kmutt.ac.th

ABSTRACT

We present a method to visualize gene co-expression from microarray data by plotting profiles of dissimilarity within gene-sets of biological pathways. A gene co-expression network is created by computing the correlation between each gene pair in a gene-set. We transform the networks into scale-free networks in order to calculate the dissimilarity weights that are used to create our profiles. Our approach further distinguishes between gene pairs consisting of both, one, or no statistically significant genes. We find that the shapes and density of the profiles provide useful information for identification of disease gene biomarkers. Our results provide a means of visualizing the overall distribution of gene dissimilarity for each gene-set, as well as how gene dissimilarity is linked to the mutual significance of gene pairs within a gene-set.

CCS CONCEPTS

•Applied computing →Bioinformatics; •Human-centered computing →*Information visualization;*

KEYWORDS

Gene co-expression network, dissimilarity, topological overlap, biomarker, gene-set profile, microarray, ANOVA, WGCNA

ACM Reference format:
Saila Shama, Philip Lu, Narumol Doungpan, Asawin Meechai, and Jonathan H. Chan. 2017. Gene-Set Profiles: Visualizing Dissimilarity Within Gene Co-Expression Networks for Biomarker Identification. In *Proceedings of International Symposium on Visual Information Communication and Interaction, Bangkok, Thailand, August 14-16, 2017 (VINCI'17),* 2 pages.
DOI: 10.1145/3105971.3108448

1 INTRODUCTION

1.1 Co-Expression and Gene-Set Data

DNA microarray is a high-throughput technique used to capture the expression pattern of genes for a certain phenotype. Studies on large amounts of genes, however, may lead to false positives in biomarker prediction. Gene-set-based analysis is a technique for reducing the dimensionality and noise within microarray data, which can improve the accuracy of detecting phenotype-correlated genes. By integrating gene-set data into co-expression networks via expression analysis, we can identify potential clusters of correlated genes for biomarker identification.

1.2 Gene Co-Expression Network

A gene co-expression network (GCN) is a fully connected graph in which each vertex corresponds to a gene and each edge weight corresponds to the correlation between a pair of genes. In this work, we construct GCNs for each gene-set, a set of genes representing a biological pathway, from the microarray data. Due to the number of nodes and connections in a fully-connected network however, it is a challenge to visualize and analyze trends in gene co-expression data. Graph visualizations, as shown in Fig. 1, are often ineffective for 2-D viewing. In this work, we propose an approach to analyzing GCNs for biomarker identification by visualizing the distribution of dissimilarity in a gene-set.

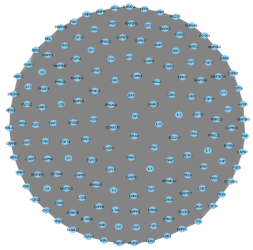

Figure 1: Visualization of GCN created using Cytoscape.

2 METHODS

2.1 Microarray Gene Expression Data

Lung cancer gene expression data (GSE10072) was downloaded from the online Gene Expression Omnibus (GEO) database [1]. The GSE10072 dataset is composed of a total of 107 samples, of which 58 are adenocarcinoma samples and 49 are non-tumor samples used as control [2]. The samples were taken from tissue samples of adenocarcinoma paired with non-involved lung tissue from current, former and non-smokers.

VINCI'17, Bangkok, Thailand
 978-1-4503-5292-5/17/08…$15.00
DOI: 10.1145/3105971.3108448

2.2 Construction of Dissimilarity Matrix

For each gene-set, the WGCNA package in R was used to create and transform the correlation matrix to a scale-free topology [3]. The topological overlap measure was then determined, where a_{ij} is the weighted expression correlation between gene i and gene j:

$$\omega_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min\left\{\sum_u a_{iu}, \sum_u a_{uj}\right\} + 1 - a_{ij}} \quad (1)$$

By subtracting equation (1) from unity, we determined the dissimilarity weight or measure for each gene pair [4].

$$d_{ij} = 1 - \omega_{ij} \quad (2)$$

The dissimilarity measure for genes was further specified by the statistical significance of each gene for all samples. Each gene was labelled as either *significant* ("sig") or *non-significant* ("non"). This was determined using the analysis of variance (ANOVA) statistical tests with a p-value of 0.05.

3 GENE-SET PROFILES

3.1 Dissimilarity Distributions

Profiles of each gene-set were created by plotting dissimilarity weights for all pairs in the set. After labelling each gene from the ANOVA tests, each gene pair was categorized as "sig-sig", "sig-non", "non-non" and assigned colors blue, green, and red, respectively, on the plots. Each profile shows the relative proportions of dissimilarity weights and how dissimilarity varies across the gene-set. For instance, the profile for the Apoptosis Signalling Pathway in Fig. 2a is skewed towards complete dissimilarity, while the Pyruvate Pathway in Fig. 3a shows a concentration of moderate dissimilarity. Profile shape may depend on the number of gene pairs in each gene-set; we find larger gene sets skew towards high dissimilarity. Generally, heavy dissimilarity within a gene-set reveals more potential biomarkers while lower dissimilarity, as with the Pyruvate gene-set, suggests fewer relevant genes. These findings can guide the use of subsequent analysis techniques such as functional modules and sub-network clustering [4].

3.2 Dissimilarity and Mutual Significance of Gene Pair

A second profile type was created, as seen in Fig. 2b and Fig. 3b, that visualizes dissimilarity split into the three possible gene pair types. The profile for the Pyruvate Pathway in Fig. 3b suggests that the co-expression of two significant genes has slightly lower dissimilarity measure compared to co-expression involving non-significant genes. These findings can validate the use of ANOVA testing for statistical significance in large samples to help analyze which genes may tend towards higher co-expression dissimilarity.

4 CONCLUSIONS

Gene-set profiles are a novel way of visualizing co-expression data to guide identification of disease gene biomarkers. Visualization of the distribution of dissimilarities over the three categories of gene pairs can validate the selection of clustering methods for subsequent analysis and statistical testing in large samples.

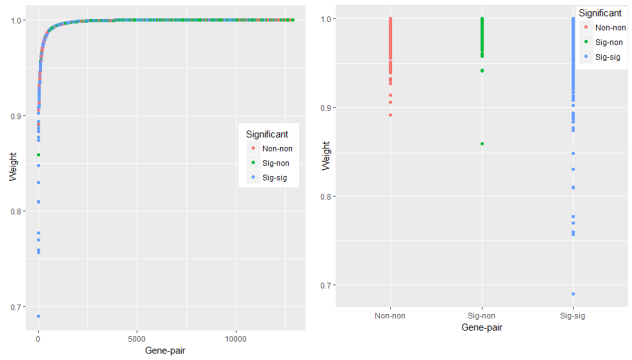

Figure 2: Gene-set profile for the Apoptosis Signaling Pathway. There are over 12000 genes in the set. (a) Plot of overall distribution shows heavy dissimilarity across the set. (b) Plot by pair category shows dissimilarity across all types.

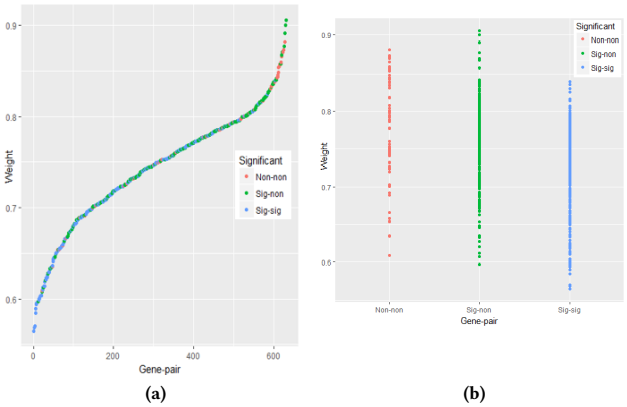

Figure 3: Gene-set profile for the Pyruvate Pathway. There are over 500 genes in the set. (a) Plot of overall distribution shows moderate dissimilarity across the set. (b) Plot by pair category shows "sig-sig" pairs having lower dissimilarity than "sig-non" or "non-non" gene pairs.

REFERENCES

[1] Tanya Barrett, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi, and Ron Edgar. 2005. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Research* 33, suppl 1 (2005), D562. DOI: http://dx.doi.org/10.1093/nar/gki022

[2] Maria Teresa Landi, Tatiana Dracheva, Melissa Rotunno, Jonine D. Figueroa, Huaitian Liu, Abhijit Dasgupta, Felecia E. Mann, Junya Fukuoka, Megan Hames, Andrew W. Bergen, Sharon E. Murphy, Ping Yang, Angela C. Pesatori, Dario Consonni, Pier Alberto Bertazzi, Sholom Wacholder, Joanna H. Shih, Neil E. Caporaso, and Jin Jen. 2008. Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLOS ONE* 3, 2 (02 2008), 1–8. DOI: http://dx.doi.org/10.1371/journal.pone.0001651

[3] Peter Langfelder and Steve Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 1 (2008), 559. DOI: http://dx.doi.org/10.1186/1471-2105-9-559

[4] Santitham Prom-on, Atthawut Chanthaphan, Jonathan H. Chan, and Asawin Meechai. 2011. Enhancing Biological Relevance of a Weighted Gene Co-Expression Network for Functional Module Identification. *Journal of Bioinformatics and Computational Biology* 9, 1 (2011), 111–119. DOI: http://dx.doi.org/10.1142/S0219720011005252