****

SCHOOL OF COMPUTING

CS5228 – KNOWLEDGE DISCOVERY AND DATA MINING

SEMESTER 1, ACADEMIC YEAR 2020/2021

FINAL REPORT

**Exploring Gene Expression Data Using**

**Weighted Gene Co-expression Network Analysis**

**Across Multiple Cancer Types**

Group Members

A.V. AKILA RAVIHANSA PERERA – A0212216X

TRAN KHANH HUNG – A0212253W

Abstract

Weighted gene correlation network analysis (WGCNA) is a powerful method that uses a topological overlap module approach for constructing co-expression networks based on gene expression data. This method involves reconstructing gene co-expression modules and summarizing modules using module eigengenes (ME) and intramodular hub genes. This enables cancer researchers to find patterns that exist among genes across multiple disparate datasets [1]. In this study, we perform a systematic assessment of differentially expressed genes in three cancer types (BRCA, GBM, OV) using data from TCGA (The Cancer Genome Atlas). The goal of the project was to construct WGCNA modules from the TCGA cancer datasets and identify the gene modules conserved across gene expression networks from different cancer types to find biologically interesting patterns. The gene modules from three gene-expression networks were subjected to multiple pairwise comparisons using Fisher’s exact test to find biologically significant gene module pairs. The gene modules that are highly significant were further analyzed using functional enrichment analysis for annotation, visualization and biological interpretation.

Table of Contents

[1.0 Introduction 3](#_Toc55997406)

[1.1 Problem Definition 3](#_Toc55997407)

[2.0 Dataset 3](#_Toc55997408)

[3.0 Methodology and Analysis 4](#_Toc55997409)

[4.0 Discussion 9](#_Toc55997410)

[5.0 References 10](#_Toc55997411)

[Appendix 11](#_Toc55997412)

# Introduction

Biological networks such as protein-protein interaction networks, metabolic networks, gene co-expression networks can be used to extract and understand system-level properties [1] and infer information about biologically important properties. Gene co-expression networks can be used as an exploratory tool to analyze system-level gene functionality by classifying differentially expressed genes into separate groups. A gene co-expression network consists of gene expression profiles that are represented by nodes in the network and significance of co-expression relationship by edges among these gene nodes. The expression profiles of these groups can be further identified by dimension reduction techniques such as principal component (PCA) and clustering analysis [2]. Although PCA is an extremely useful tool for identifying features that reveals relative similarity among the samples it is not effective for high dimensional gene expression data because it is not capable of providing any reasons or ways to further investigate the differences between sample types [3].

WGCNA (Weighted Gene Co-expression Network Analysis) is a general framework for analyzing gene co-expression networks to find related patterns in protein interactions which can provide insights into the underlying cellular processes [4]. It employs a node dissimilarity measure that considers topological similarity in the network as a whole to identify network modules (clusters of nodes). These gene modules have similar expression profiles and may also have interconnected biological functions. A module eigengene is a vector that represents gene expression profile in a module and also it is the first principal component of the co-expression module [5]. By focusing on a network analysis that takes a global approach instead of focusing on individual genes, WGCNA method can reveal system-level mechanism of cancer related genes and identify potential biomarkers that can lead to the development new drugs.

## Problem Definition

In this study we focus on three highly prevalent cancers in humans that affects millions of people today. We aim to investigate the common properties of genes in the gene co-expression networks of three cancer types; Glioblastoma Multiforme (GBM), Ovarian Serous Cystadenocarcinoma (OV), and Breast Invasive Carcinoma (BRCA). We’ve focused on three primary questions;

* Are there network properties that are conserved across these cancer types?
* Are there highly significant overlapping gene modules across cancer types?
* What are biological functional similarities and/or dissimilarities among overlapping gene modules?

# Dataset

All the datasets are already harmonized using the same genomic data processing pipeline in TCGA and mean-normalized across all of TCGA data [6] [7] [8]. Therefore, no further preprocessing is required.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Genes** | **Samples** |
| TCGA glioblastoma multiforme (GBM) gene expression by RNAseq | 20,531 | 172 |
| TCGA ovarian serous cystadenocarcinoma (OV) gene expression by RNAseq | 20,531 | 308 |
| TCGA breast invasive carcinoma (BRCA) gene expression by RNAseq | 20,531 | 1218 |

Table 1. Description of input dataset and method of extraction.

# Methodology and Analysis

A row in a dataset contains the differential expression ratio of a gene for each sample. First, we filter the dataset for obsolete genes (marked withdrawn or pseudo genes) and select the genes with expressional variance ranked in the top 75 percentile of each dataset. This is because in general there is a percentage of genes that are not expressed and we need remove values that are considered noise and less reliable which will make the resulting gene network more robust. We are looking for genes that have higher variance that might have caused due to biologically driven mutations.

![Diagram

Description automatically generated]()

Figure 1. Project analysis workflow based on WGCNA framework.

Gene expression network is defined using a gene co-expression similarity measure; which denotes absolute value of the Pearson correlation between gene pair *i* and *j*. An adjacency function then transforms the similarity matrix into an adjacency matrix that contains connection strengths. We use a soft thresholding function [4] that helps to reduce noise while not compromising the overall connectivity of the network. A hard threshold might result in a loss of co-expression network due to binary encoding fashion.

**Selecting a soft threshold power term**

The lowest possible should be selected that leads to an approximately scale-free network [5]. A scale-free network in general consists of large hubs (highly connected central nodes) while maximizing the overall mean connectivity. We aim to conform to a scale-free network because most biological networks follow this power-law degree distribution [9] and our goal is to build a biologically realistic network. The figure 2 illustrates the scale free topology model fit for a range of soft threshold values. As a general rule we pick the lowest possible power term where topology approximately fits a scale free network while maximizing the overall mean connectivity.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Figure 2. Scale-free Topology Model Fit and Mean Connectivity measure for a range of soft thresholds.

We selected soft thresholds values 6, 9, 3 for BRCA, GBM, OV cancer type datasets respectively. The next step after network construction is module detection. Modules are groups of closely interconnected nodes (prominent subnetworks within the main network). The WGCNA package in R uses an unsupervised clustering method to identify gene modules. The identified gene modules are depicted in figure 4. The colour row in the bottom indicates the assigned module. The grey bars indicate genes that were not assigned to any module. The eigengene dendrogram and the eigengene adjacency heatmap (in figure 3) represents tight clusters of correlated eigengenes. The eigengene can be considered as a weighted average expression profile of a gene module. It can be shown that hub genes (i.e. most highly connected gene within a module) are highly correlated with the module eigengene [5]. According to eigengene adjacency heatmap in figure 3 it can be seen that there are clusters of highly correlated eigengenes along the diagonal. These clusters were further analyzed by comparing against gene modules in other cancer types in a pairwise fashion.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Figure 3. Eigengene adjacency heatmap and eigengene clustering dendrogram.

In WGCNA package the modules are defined as branches of the tree and a height cut-off threshold can be used to control robustness of the modules; i.e. small cut-off values lead to tightly connected genes and larger values lead to big modules. However, this method is not good at detecting genes that are in the borderline of the tree cut-off level. A dynamic branch cutting method [10] was introduced by the authors to circumvent the shortcomings of static height cut-off method in which the surrounding branches are explored and tested for a given set of criteria to determine whether nearest clusters are close enough to be merged. For the purpose of this study, we use dynamic cut tree method with following parameter values.

* deepSplit: 4 (highest sensitive)

This parameter controls the sensitivity of module detection for module splitting. We had to fine tune this parameter (increase from default 2) because our dataset resulted in large modules with more than 2000 genes. But this is problematic when performing functional gene enrichment analysis since most tools do not support large gene sets partly because it is hard to interpret enrichment results from large gene sets. Furthermore, by influencing dynamic cut tree algorithm to find more tightly connected gene networks we are able to identify common biological functions that are shared across these cancer types (when performing pairwise gene module testing).

* minClusterSize: 30 (minimum number of genes needed to be considered as a module)

Increasing this value will cause the total number of modules to decrease. It also helps to avoid generating similar clusters of few genes by forcing them to be merged into one. We have selected the default minimum cluster size as 30 since it is sufficient to run a gene enrichment analysis.

A network topological overlap heatmaps were produced from TOM-based dissimilarity measures for each cancer type. The resulting TOM plots with dendrogram and module membership colors are shown in figure XX (in Appendix). Each row and column corresponds to a gene. Light colors represents low topological overlap and dark red represents high topological overlap. Dark red squares along the diagonal represents gene modules.

**Pairwise Gene Module Comparison**

Our aim is to find biologically interesting patterns that underlie across different cancer types. To that end, we’ve performed a pairwise comparison of gene modules from each network to identify biologically significant gene modules. The *p* value of the Fisher's exact test [11] is used to test the null hypothesis which is a test for independence as opposed to association. We test whether genes in a gene module pair happen to be there by random chance or they are statistically related. The null hypothesis is usually assumed to have no relationship between the two objects, so if the *p* value is less than 0.05, the null hypothesis can be overturned, and the opposite hypothesis is derived. In the experiment, the *p* value of every module in one cancer type is calculated with each module in another cancer type in a pairwise fashion. When the result *p* value is less than 0.05, the two modules are related.

**Gene Enrichment and Pathway Analysis**

Pathway analysis identifies the specific protein functions, biological pathways, and physical interactions that are enriched in a particular group. It involves calculating over and under representation of a gene set against a background dataset. This reference data are obtained from well-known standard data sources such as GO (Gene Ontology), KEGG, and STRING. Pathway enrichment analysis helps researchers gain insights into the underlying mechanism of how these genes interact together to perform various biological functions. In this study, we’ve used STRING online tool [12] to obtain three classes of biological functions; cellular components, molecular function, and biological process. We compare the results from pathway analysis using the following method

* Perform pairwise comparison of gene modules across cancer types and find the gene module pair with the lowest *p* value (gene sets least likely to have occurred due to chance)
* Find the intersection of gene modules for all such pairs (*BRCA & GBM*, *GBM & OV*, *OV & BRCA*) and run gene enrichment and pathway analysis using STRING tool
* Repeat the same for genes that are unique to each gene module in a pair (do this for all cancer type pairs)
* Find biological functions that are conserved or not conserved across cancer types
* Find biological functions that are unique in these gene modules
* Draw conclusions about the impact of cancers on biological processes.

|  |  |
| --- | --- |
| GBM | GBM |
| GBM | GBM |
| GBM | GBM |

Figure 4. Clustering gene dendrogram and module eigengene (ME) dendrogram.

The results from pairwise comparison of gene modules are shown in figure XX (in Appendix). The heatmaps show the relationship between gene module pairs across cancer types. White color cells indicate gene modules having the highest statistical significance *(pvalue ~ 0, differences in characteristics are high in the population).* The most significant gene module pair selected is indicated using a yellow diamond and its composition is shown in a Venn diagram. The highest overlap of genes were recorded in *OV & BRCA* pair of 707 genes while overlap of *GBM & OV*, *BRCA & GBM* was found to be 381 and 556 respectively.

# Discussion

# References

|  |  |
| --- | --- |
| [1] | F. LI, "Human diseases through the lens of network biology," *Trends Genet,* pp. 150-9, 2012. |
| [2] | P. Langfelder, T. Fuller, J. Dong, A. Li and S. Hovarth, "Weighted gene coexpression network analysis: state of the art," *Journal of biopharmaceutical statistics,* vol. 20, no. 2, pp. 281-300, 2010. |
| [3] | M. DiLeo, G. Strahan, M. den Bakker and O. A. Hoekenga, "Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome.," *PLoS One,* vol. 6, no. 10, 2011. |
| [4] | B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis.," *Statistical applications in genetics and molecular biology,* vol. 4, 2005. |
| [5] | P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics,* vol. 9, no. 1, 2008. |
| [6] | "OV dataset: gene expression RNAseq," 2017. [Online]. Available: https://tcga.xenahubs.net/download/TCGA.OV.sampleMap/HiSeqV2\_PANCAN.gz. [Accessed 1 10 2020]. |
| [7] | "GBM dataset: gene expression RNAseq," 2017. [Online]. Available: https://tcga.xenahubs.net/download/TCGA.GBM.sampleMap/HiSeqV2\_PANCAN.gz. [Accessed 1 10 2020]. |
| [8] | "BRCA dataset: gene expression RNAseq," 2017. [Online]. Available: https://tcga.xenahubs.net/download/TCGA.BRCA.sampleMap/HiSeqV2\_PANCAN.gz. [Accessed 1 10 2020]. |
| [9] | R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science,* vol. 118, no. 21, pp. 4947-4957, 2005. |
| [10] | B. Zhang and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.," *Bioinformatics,* vol. 24, no. 5, pp. 719-720, 2008. |
| [11] | R. A. S. Fisher, Statistical methods for research workers, Edinburgh: Oliver and Boyd, 1934, pp. 524-525. |
| [12] | "STRING: functional protein association networks," [Online]. Available: https://string-db.org/. [Accessed 1 10 2020]. |

# Appendix

|  |
| --- |
|  |
|  |
|  |
| Figure XX. Sample clustering dendrogram for each cancer type. |

|  |
| --- |
|  |
|  |
|  |
| Figure XX. Network heatmap of all genes for each cancer type. |

|  |  |
| --- | --- |
| BRCA_GBM |  |
| GBM_OV | GBM_OV |
| OV_BRCA | OV_BRCA |
| Figure XX. Heatmaps of pairwise gene modules (left side) and the gene composition venn diagram of the most significant gene module pair (right side). The most significant pair identified is marked as a diamond in yellow color in the heatmap. | |