CONSTRUCTION OF GRN FOR LUNG CANCER DATASETS USING WEIGHTED CO-EXPRESSION NETWORKS

SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

H. M. RAFSANZANI
STUDENT ID. 144429
SAMIN ANAN
STUDENT ID. 144444

UNDER THE SUPEVISION OF

TAREQUE MOHMUD CHOWDHURY
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (CSE)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
NOVEMBER 2018

CONSTRUCTION OF GRN FOR LUNG CANCER DATASETS USING WEIGHTED CO EXPRESSION NETWORK

Supervised By

Tareque Mohmud Chowdhury
Assistant Professor, Department of CSE
Islamic University of Technology (IUT)

Prepared By

H. M. Rafsanzani (144429)
Samin Anan (144444

Declaration of Authorship

This is to certify that the work presented in this these is the outcome of the analysis and experiments carried out by **H. M. Rafsanzani** and **Samin Anan** under the supervision of **Tareque Mohmud Chowdhury**, Assistant Professor of Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

- -
H. M. Rafsanzani
Student ID – 144429

- -
Samin Anan
Student ID – 144444

Supervisor:

- -
Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

Contents

Abstract

Gene co-expression networks can be used to associate genes of unknown function with biological processes, to prioritize candidate disease genes or to discern transcriptional regulatory programmes. With recent advances in transcriptomics and next-generation sequencing, co-expression networks constructed from RNA sequencing data also enable the inference of functions and disease associations for non-coding genes and splice variants. Although gene co-expression networks typically do not provide information about causality, emerging methods for differential co-expression analysis are enabling the identification of regulatory genes underlying various phenotypes. Correlation networks are increasingly being used in bioinformatics applications. For example, weighted gene co-expression network analysis is a systems biology method for describing the correlation patterns among genes across microarray samples. Weighted correlation network analysis (WGCNA) can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module eigengene or an intramodular hub gene, for relating modules to one another and to external sample traits (using eigengene network methodology), and for calculating module membership measures. Correlation networks facilitate network based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. Gene regulatory networks can be used to identify the genes of cancer affected patients that are responsible for tumor formation. We provide a method to use weighted gene co-expression network analysis to identify genes that are responsible for cancer patient based on clinical trait information by cross-matching with healthy patient genes.

1. Introduction

1.1 Gene Regulatory Network

A gene (or genetic) regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. These play a central role in morphogenesis, the creation of body structures, which in turn is central to evolutionary developmental biology.

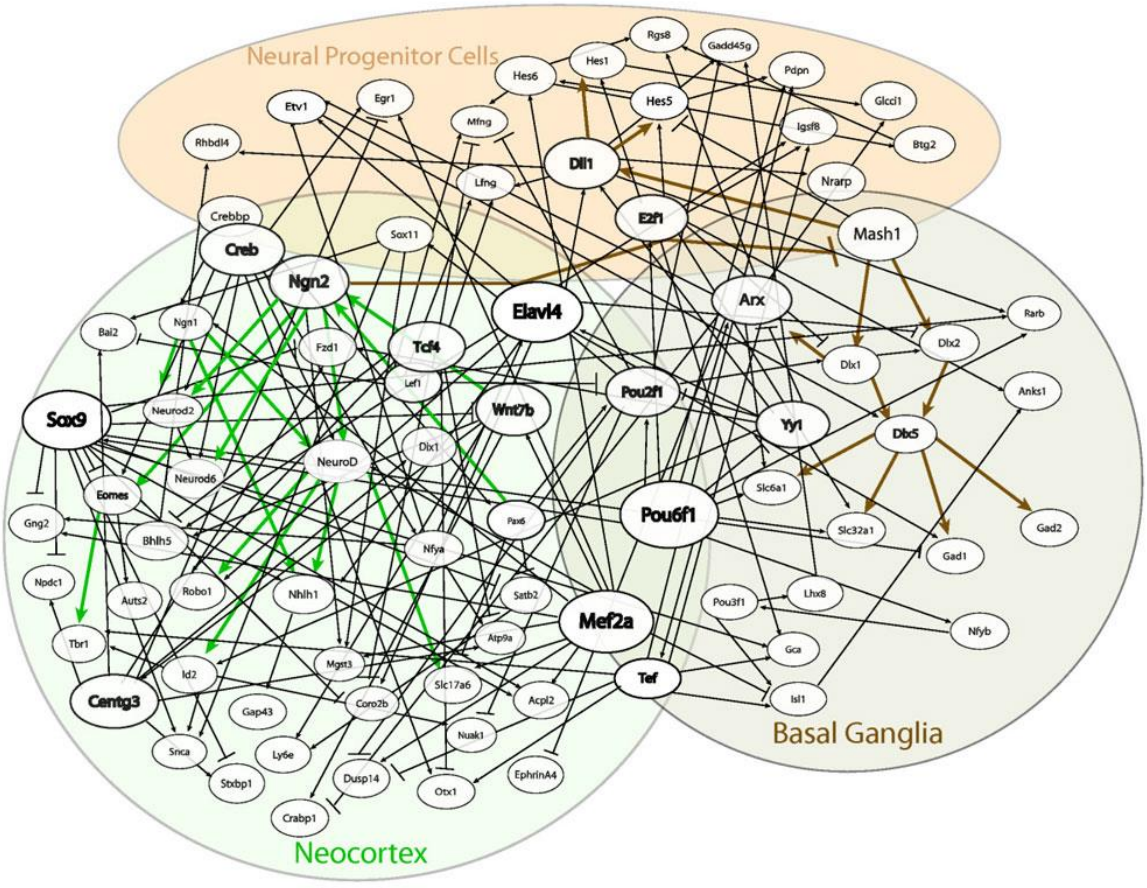

Figure: A gene regulatory network to differentiate between Basal Ganglia, Neocortex and Neural Progenitor cells

1.2 Lung Cancer

Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale.

Lung cancer is the leading cause of cancer deaths in the United States, among both men and women. Lung cancer claims more lives each year than do colon, prostate, ovarian and breast cancers combined.

People who smoke have the greatest risk of lung cancer, though lung cancer can also occur in people who have never smoked. The risk of lung cancer increases with the length of time and number of cigarettes you've smoked. If you quit smoking, even after smoking for many years, you can significantly reduce your chances of developing lung cancer.

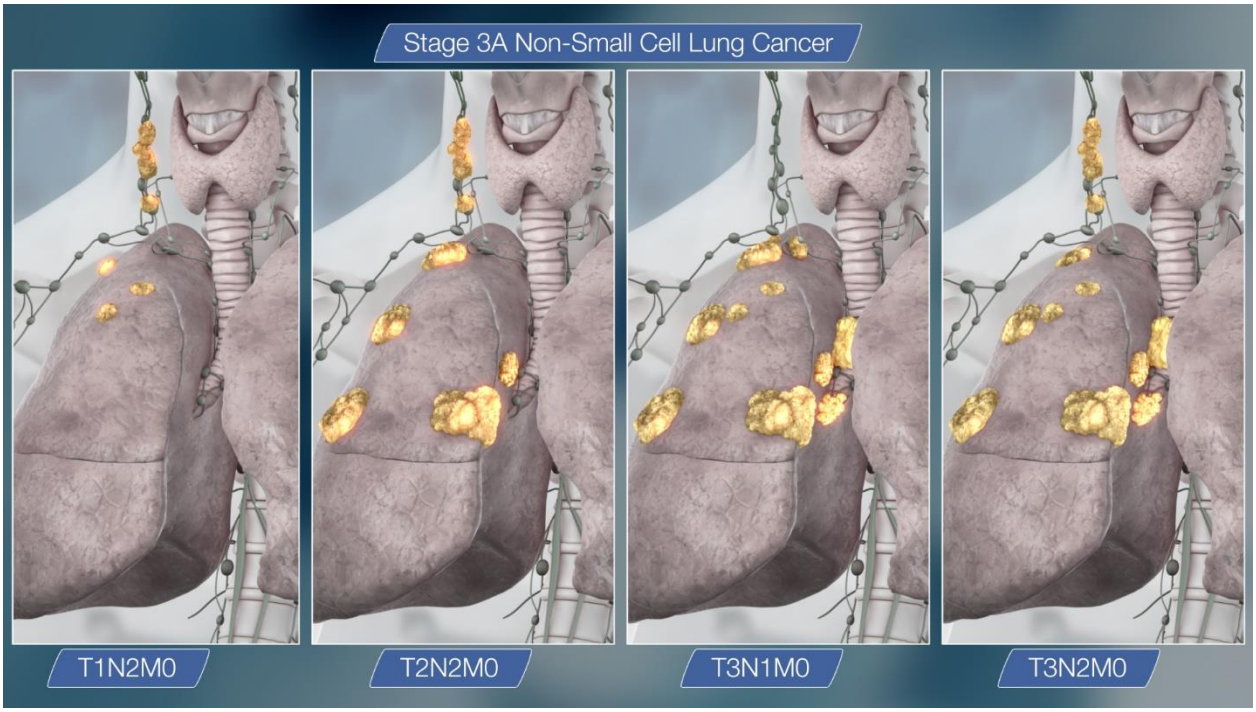

Figure: Lung cancer stage 4.

1.3 Weighted correlation network

Weighted correlation network, also known as weighted gene co-expression network (WGCNA), is a widely used data mining method especially for studying biological networks based on pairwise correlations between variables. While it can be applied to most high-dimensional data sets, it has been most widely used in genomic applications. It allows one to define modules (clusters), intramodular hubs, and network nodes with regard to module membership, to study the relationships between co-expression modules, and to compare the network topology of different networks.

The application of complex network modeling to analyze large co-expression data sets has gained traction during the last decade. In particular, the use of the weighted gene co-expression network analysis framework has allowed an unbiased and systems-level investigation of genotype-phenotype relationships in a wide range of systems

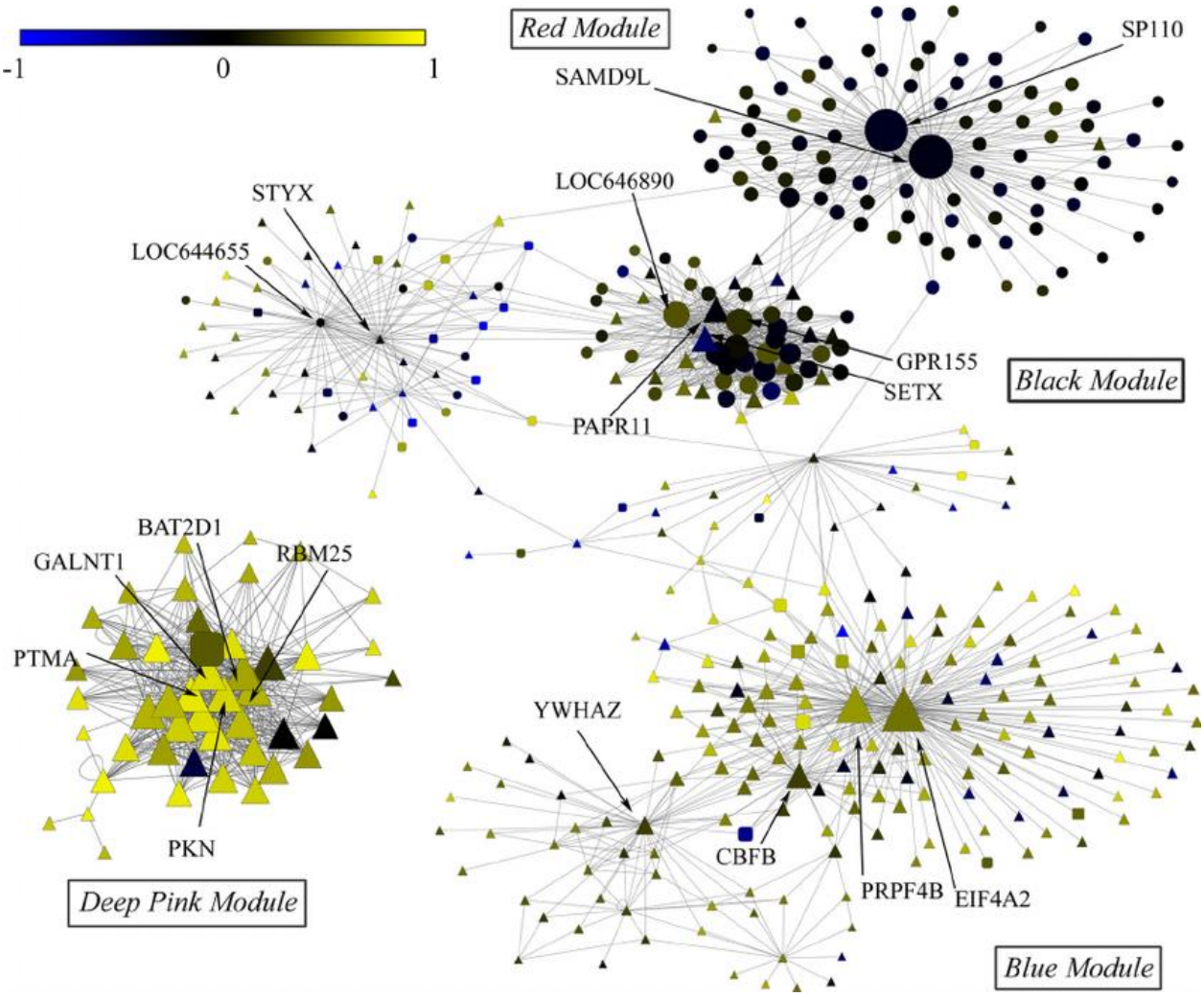

Figure: Some sample gene regulatory networks.

1.4 Weighted Correlation Network Analysis (WGCNA)

Weighted correlation network analysis (WGCNA) can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module Eigen gene or an intramodular hub gene, for relating modules to one another and to external sample traits (using Eigen gene network methodology), and for calculating module membership measures. Correlation networks facilitate network based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, e.g. cancer, mouse genetics, yeast genetics, and analysis of brain imaging data.

WGCNA has been widely used for analyzing gene expression data (i.e. transcriptional data), e.g. to find intramodular hub genes.

It is often used as data reduction step in systems genetic applications where modules are represented by "module Eigen genes". Module Eigen genes can be used to correlate modules with clinical traits. Eigen gene networks are co-expression networks between module Eigen genes (i.e. networks whose nodes are modules). WGCNA is widely used in neuroscientific applications, and for analyzing genomic data including microarray data, single cell RNA-Seq data. DNA methylation data, miRNA data, peptide counts and microbiota data (16S rRNA gene sequencing). Other applications include brain imaging data, e.g. functional MRI data.

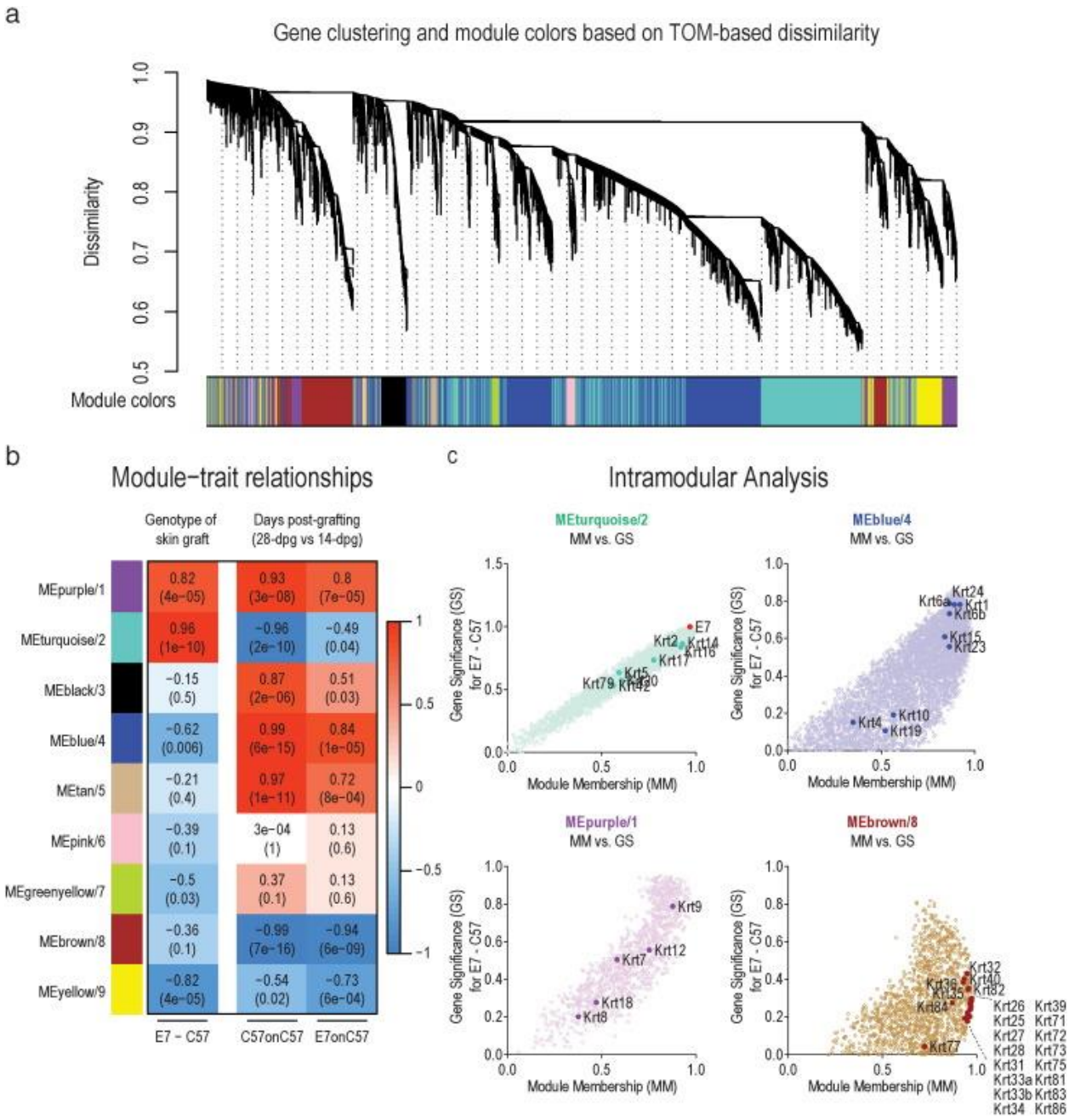

Figure: a) Hierarchical clustering of a GRN b) The module-trait relationship between gene expression dataset and trait dataset c) Intramodular analysis

2. Synthesis of GRN

Solving a GRN related to any development process requires knowledge of what transcription factors and signal molecules are involved in the system, when and where the genes are expressed, and most importantly, how the regulatory genes interact with one another. None of these data is simply acquired, and extensive measurements are typically needed to characterize the temporal and spatial expression of all relevant regulatory genes, while a huge number of data points is required to identify the genuine linkages among interacting genes as the activities of individual genes are alternately removed from the system one by one.

The first step in building a GRN is to identify the regulatory genes involved. When the complete genome sequence is available, the most comprehensive and arguably the best solution is a top-down approach, i.e., a genome-wide survey of all predicted regulatory genes followed by characterization of their spatial and temporal expression patterns.

3. Problem Domain and Problem Statement

3.1 Challenges & Research Issues

There are some challenges in this problem domain. Those are discussed below –

- The dataset for lung cancer has not been well built and correlation between data had to be done.
- There was insufficient trait information available for the healthy patients which caused a data shift in the final result.
- The gene expression data and trait data was in incompatible format.
- The library functions in R “GO.db” packages in the recent update was flawed to the point that some library functions had to be re-written to make the code work.

3.2 Problem Statement

- To construct a gene regulatory network from Merged Lung Cancer Datasets
- To compare cancer patient GRN with healthy patient GRN
- Merging multiple existing Lung cancer datasets
- Use trait module information to track module of genes responsible for cancer.
- Using the module trait relation finding the genes responsible for lung cancer.

4. Existing Methods to Form GRN

There are some existing methods to form GRN. Some of them are given below.

- WGCNA: A tool that constructs a co-expression network using Pearson correlation (default) or a custom distance measure.
- ARCANE: A tool that removes indirect connections between genes (i.e. partners of a gene that have a stronger correlation with each other than with the gene itself), leaving only those connections that are expected to be regulatory.
- GENIE3: A tool that incorporates TF information to construct a regulatory network by determining the TF expression pattern that best explains the expression of each of their target genes.
- Combat: A method that is robust to outliers and also effective at batch effect correction in small sample sizes (<25).
- Dicer: A method that identifies modules that correlate differently between sample groups, e.g. modules that form one large interconnected module in one group compared with several smaller modules in another group.
- DiffCoEx: A method that uses a similar approach to WGCNA to identify and group differentially co-expressed genes instead of identifying co-expressed modules.
- Dingo: DINGO is a more recent tool that groups genes based on how differently they behave in a particular subset of samples (representing e.g. a particular condition) from the baseline co-expression determined from all samples

- HO-GSVD: A tool similar to GSVD, but that can be used across multiple sample groups rather than only two.
- David: A widely used tool with an online web interface. Users supply a list of genes and select the annotation categories from various sources to identify enrichment.
- g:Profiler: A tool that performs enrichment analyses for gene ontologies, KEGG pathways, protein– protein interactions, TF and miRNA binding sites.
- GIANT: Tissue-specific interaction network database. Includes 987 Datasets encompassing 38 000 conditions describing 144 tissues types.

5. Literature Review

Title: E-MTAB-6043 - A microarray meta-dataset of non-small cell lung cancer

Conference: European Bioinformatics Institute 23 March 2018

- 41 different types of methods are described
- Comparison between different kinds of existing methods and enlisting features
- Creating a benchmark for module creation methods
- Popular methods such as WGCNA, Dicer, DiffCoEx, GSCNA are described with benefits.

Title: WGCNA: An R package for weighted correlation network analysis

Article: bmcbioinformatics 29 December 2008

- Analysis of the WGCNA package and helpful R function related to it.
- Sample construction of Mouse Liver gene expression data and comparison of it.
- Describes the ins and outs of the Weighted gene co-expression analysis method.

Title: WGCNA: An R package for weighted correlation network analysis

Article: bmcbioinformatics 29 December 2008

- Analysis of the WGCNA package and helpful R function related to it.
- Sample construction of Mouse Liver gene expression data and comparison of it.
- Describes the ins and outs of the Weighted gene co-expression analysis method.

Title: A merged lung cancer transcriptome dataset for clinical predictive modeling

Article: nature.com, 24 July 2018

- 17 different datasets merged into different cluster of datasets
- Forms world's largest (till date) lung cancer dataset information
- Published and updated from 2009 till date.

6. Datasets

1	Character	Age	Overall_st	Overall_st	TNMstage	TNMstage	NewSmok	NewGend	NewRecu	NewStage	NewHistology
2	GSM12136	63.57	49.32	1	1	0	3	1	0	1	1
3	GSM12136	68.15	90.24	0	1	0	3	0	0	1	1
4	GSM12136	69.15	87.84	0	2	0	3	1	0	2	1
5	GSM12136	82.36	82.08	0	1	0	2	0	0	1	1
6	GSM12136	68	22.92	1	2	0	3	1	1	2	1
7	GSM12136	57.31	84.36	0	2	0	3	1	0	2	2
8	GSM12136	63.63	82.2	0	2	1	2	0	0	4	2
9	GSM12136	71.43	51.96	1	2	0	2	1	1	2	1
10	GSM12136	44.21	74.64	0	2	0	1	0	0	2	1
11	GSM12136	74.16	48.96	1	2	1	1	0	1	4	1
12	GSM12136	66.46	75.72	0	1	0	2	1	0	1	1
13	GSM12136	59.01	71.88	0	2	0	2	0	0	2	1
14	GSM12136	70.56	63.84	0	2	0	2	0	0	2	1
15	GSM12136	72.96	66.72	0	1	0	2	0	0	1	1
16	GSM12136	79.28	60.24	0	2	0	2	1	0	2	1
17	GSM12136	73.26	41.64	0	2	1	2	0	0	4	2
18	GSM12136	52.14	81.12	0	1	0	3	1	0	1	1
19	GSM12136	85.78	34.08	1	2	0	2	1	0	2	1
20	GSM12136	76.93	77.52	0	2	0	3	1	0	2	1
21	GSM12136	73.23	69.12	0	1	0	2	1	0	1	1
22	GSM12136	77.18	59.4	0	1	0	1	0	0	1	1
23	GSM12136	78.96	61.44	0	2	1	2	0	1	4	1
24	GSM12136	78.99	76.68	0	2	0	3	1	0	2	1
25	GSM12136	55.39	68.88	1	1	0	1	0	0	1	1

Figure: Merged trait dataset of cancer-affected and heralthy patients.

1	Tr	loc	ge
2	8039748	1	A1BG
3	7960947	2	A2M
4	7959786	65985	AACS
5	8083415	13	AADAC
6	8103706	51166	AADAT
7	8052798	22848	AAK1
8	8002347	16	AARS
9	7943552	60496	AASDHPPT
10	8142554	10157	AASS
11	7993126	18	ABAT
12	8162940	19	ABCA1
13	8058708	26154	ABCA12
14	8132743	154664	ABCA13
15	7998784	21	ABCA3
16	7917798	24	ABCA4
17	8018038	23461	ABCA5
18	8017964	23460	ABCA6
19	8024120	10347	ABCA7
20	8017885	10351	ABCA8
21	8017927	10350	ABCA9
22	8140782	5243	ABCB1
23	7924956	23456	ABCB10
24	8140752	5244	ABCB4
25	8059111	10058	ABCB6

Figure: Gene annotation information

1	Genes	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€	GSM1213€
2	1	A1BG	6.193788	5.120598	5.402096	6.148757	5.764026	6.132949	6.129625	5.063807	6.234404	5.614969	5.828414	5.955376	5.451649	6.007381	5.352719	5.548329	5.119513	5.237179
3	2	A2M	11.45692	11.84258	13.04357	12.16883	12.69036	11.75635	9.750002	9.193014	10.57331	12.64902	13.08817	10.84209	12.32176	10.93218	12.5653	10.76255	13.06907	11.89512
4	3	AACS	8.199319	7.932785	7.206544	7.597157	7.843723	7.360108	9.532414	7.78836	8.45423	7.84192	7.4762	8.070317	9.061497	8.367668	8.128292	9.020604	7.870464	8.578635
5	4	AADAC	7.532517	4.931691	4.72009	5.225023	6.476748	6.06874	8.589095	4.730841	8.655138	4.880604	7.824951	6.711232	4.719891	4.811638	6.182722	5.069924	4.629495	5.540632
6	5	AADAT	6.245118	5.697312	5.714073	5.493415	6.050574	5.129941	5.334876	5.508959	6.227409	5.621714	5.486997	6.197936	5.547385	5.647356	6.180448	6.397452	6.507356	5.618148
7	6	AAK1	9.770302	9.095166	8.87376	9.024537	9.452556	8.965902	9.176045	8.733507	9.200983	8.660739	8.627437	8.515107	8.145968	8.54221	8.954914	8.589328	8.165147	8.44766
8	7	AARS	9.377016	8.680404	8.734925	9.08997	9.754274	9.600239	9.983395	8.94059	10.11381	9.978111	9.072017	10.02218	9.952385	9.102123	9.321562	9.760724	9.063483	9.3948
9	8	AASDHPP	8.745791	9.415227	9.301868	9.380452	8.906762	9.31634	8.997385	9.250964	9.313895	9.76191	9.29508	9.705873	8.833771	8.914336	9.495678	9.769729	9.83649	9.130769
10	9	AASS	5.822601	6.145239	7.223604	7.027268	6.5276	5.559956	5.494396	7.374195	6.209636	6.20045	5.682973	7.222758	5.185173	5.354856	6.387179	4.938062	7.071207	4.919488
11	10	ABAT	6.466821	7.041783	8.692121	6.74227	6.619884	6.791062	6.664212	6.489958	6.612111	8.016339	7.62922	6.480246	7.352177	6.686875	6.768217	5.83458	7.471374	7.36569
12	11	ABCA1	9.018026	10.12783	9.876599	9.080118	9.473704	9.578843	8.934476	8.59911	8.120552	9.391065	9.376795	9.966338	7.324492	9.232959	8.605225	8.394182	9.015424	9.814832
13	12	ABCA12	4.293317	4.595655	3.487759	3.719693	3.551182	3.822482	4.188296	3.373815	4.440461	5.192124	3.284603	5.312359	3.823744	3.681243	3.45384	3.875232	4.89545	3.852253
14	13	ABCA13	7.597453	3.314835	7.192179	3.14651	3.230511	3.994488	5.27079	4.00416	4.690136	3.373813	6.160208	3.43226	3.146896	3.289551	8.208515	3.149503	7.892224	5.830638
15	14	ABCA3	7.754172	8.859201	9.256076	8.162136	7.153897	6.521938	6.629163	6.354627	7.277945	9.008151	7.702471	8.912827	9.354246	8.453247	8.527472	5.955146	9.856688	9.676669
16	15	ABCA4	5.497274	5.813432	4.966972	5.444102	5.352655	4.987619	5.167841	4.772208	5.136233	4.968031	4.692477	5.357072	6.111178	7.782145	7.442401	5.008308	8.247275	5.183853
17	16	ABCA5	6.515668	7.498788	7.513148	6.885574	6.442335	7.199944	6.711946	7.15005	6.625759	6.167021	7.498043	6.731281	5.836074	6.071847	7.675266	6.448522	7.375707	6.334161
18	17	ABCA6	4.485823	6.254699	6.361564	4.561955	4.19436	6.130484	3.9911	4.209085	4.602447	4.988813	6.762571	4.707808	4.317087	4.184537	4.90099	4.117766	7.390599	4.335086
19	18	ABCA7	6.46463	6.906049	6.986304	6.816967	6.298519	6.075534	6.492797	7.780506	6.286268	7.029005	6.689441	7.296525	7.346183	7.899174	6.60753	6.220895	7.093463	6.434301
20	19	ABCA8	6.317173	8.58603	10.42445	6.356676	5.170876	8.69127	5.242135	5.269896	5.649596	5.066155	11.38616	5.507646	5.294168	5.12265	7.855288	4.855576	10.00276	6.921763
21	20	ABCA9	5.146895	5.720622	5.639488	4.172938	4.288126	6.452277	4.340652	4.181722	4.201861	4.295169	6.386843	4.236523	4.206999	4.102711	5.061895	4.270891	4.941575	4.256474
22	21	ABCB1	4.799186	5.467065	5.028215	5.151007	4.257571	5.014368	4.20883	5.195646	4.518321	4.531345	5.351985	5.803763	4.530002	5.116435	4.473445	4.77355	6.275834	4.975841
23	22	ABCB10	7.040348	8.100322	8.042354	8.280972	7.404819	7.454406	7.754939	8.118626	8.092883	8.091758	7.890731	8.239813	7.052988	8.09998	7.936622	7.960155	8.646204	8.177877
24	23	ABCB4	4.687843	4.771974	4.487615	3.88484	5.250662	5.015516	4.192255	5.306791	4.209118	4.71614	3.955768	5.096561	4.425318	4.244384	4.234351	4.300127	3.941281	6.013874
25	24	ABCB6	7.489629	6.018398	6.196719	6.586124	6.824417	6.228864	9.250914	6.938734	7.203	7.722698	6.71893	6.5182	7.867419	6.644636	6.528204	6.471683	7.041244	6.432988

Figure: Gene-expression dataset

7. Research Challenges

Three were several challenges that occurred during the execution of the idea for finding the culprit genes that caused the spur of cancer tumors. The method selected was to follow a weighted gene regulatory network analysis approach. Ion the proposed method the R studio was used to operate on the gathered datasets.

1. The dataset formed for the execution of the data had not been well built and thus it the retrieved data from the internet could not been used directly to process.
2. Hence the dataset had to be trimmed and formatted to the point that co-relation could be possible.
3. The trait information for healthy patients was not sufficient. Therefore, a shift in the final data could have affected the results a bit.
4. The gene expression data and trait data was in incompatible format.
5. The library functions in R “GO.db” packages in the recent update was flawed to the point that some library functions had to be re-written to make the code work.

8. Relative Research Extension

8.1 Integrating Gene Regulatory Networks to identify cancer-specific genes

When an organism is subjected to a different condition either internal or external to it (environmental changes, stress, cancer, etc.), its underlying mechanisms undergo some changes. To build robust and reliable Gene Regulatory Networks (GRNs) from microarrays, it is necessary to integrate multiple data collected from other researches. To identify links in common among a set of independent studies, researchers apply consensus networks analysis. For example, a clustering technique can be applied and coupled with a statistically based gene functional analysis for the identification of novel genes. Again, group genes that perform similar functions into 'modules' and then build networks of these modules to identify mechanisms at a more general (higher) level. More recently, a similar approach was applied to a large number of cancer datasets where case and control are compared. For each dataset, the pairwise correlation of gene expression profile is computed and a frequency table is built. Then the values in the table are used to build a weighted gene co-expression frequency network. After this they identify sub-networks with similar members and iteratively merge them together to generate the final network for both cancer and healthy tissue.

8.2 Weighted Frequent Gene Co-Expression Network Mining to Identify Genes Involved in Genome Stability

Distinct types of human cancer share similar traits, including rapid cell proliferation, loss of cell identity, and the ability to migrate and seed malignant tumors in distal locations. Understanding these common traits and identifying the underlying genes/networks are key to gaining insight into cancer physiology, and, ultimately, to prevent and cure cancer. With cancer gene expression microarray datasets increasingly accumulated in central repositories, many bioinformatics data analysis methods have been developed to identify cancer related genes, characterize cancer subtypes and discover gene signatures for prognosis and treatment prediction. As an example, in breast cancer research, a supervised approach was adopted to select 70 genes as biomarkers for breast cancer prognosis and was successfully tested in clinical settings. However, a major drawback of such an approach is that the selected gene features are usually not functionally related and hence, cannot reveal key biological mechanisms and processes behind different patient groups.

In order to overcome this hurdle to identify functionally related genes associated with disease development and prognosis, several approaches have been adopted. One such approach is gene co-expression analysis, which identifies groups of genes that are highly correlated in expression levels across multiple samples. The metric to measure the correlation is usually the correlation coefficient (e.g., Pearson correlation coefficient or PCC) between expression profiles of two genes. Using this approach, we were able to identify new gene functions in regulating cell mitosis in breast cancer by studying genes that have high correlation with the expression of the DNA repair protein, BRCA1.

By applying an advanced network mining algorithm, dense modules of highly co-expressed genes can be identified which can lead to the discovery of new gene functions, disease genes and biomarkers. For example, Horvath's group has developed a series of weighted gene co-expression network analyses using a hierarchical clustering based approach [6], [10], [12]– [15]. This method was applied to identify disease-associated genes such as ASPM in glioblastoma.

8.3 Gene expression profiling predicts clinical outcome of breast cancer

Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumors according to their clinical behavior. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70-80% of patients receiving this treatment would have survived without it. None of the signatures of breast cancer gene expression reported to date allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumors of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumor cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumors of BRCA1 carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and angiogenesis. This gene expression profile will outperform all currently used clinical parameters in predicting disease outcome. Our findings provide a strategy to select patients who would benefit from adjuvant therapy.

8.4 Computational methods to dissect gene regulatory networks in cancer

Cancer is a disease of gene dysregulation, where cells acquire genetic alterations that drive aberrant signaling. These alterations adversely impact transcriptional programs and cause profound changes in gene expression. Large international consortia have generated massive tumor profiling data sets across many cancer types, collecting mutation and copy number variation, mRNA expression, and in some cases epigenomic and proteomic profiles. An overarching goal of these tumor-profiling efforts is to identify genes that are essential drivers of cellular processes in cancer. Here we review diverse computational methodologies that have sought to interpret somatic alterations and gene expression data through models of gene regulatory networks.

8.5 The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks

In this study, we infer the breast cancer gene regulatory network from gene expression data. This network is obtained from the application of the BC3Net inference algorithm to a large-scale gene expression data set consisting of 351 patient samples. In order to elucidate the functional relevance of the inferred network, we are performing a Gene Ontology (GO) analysis for its structural components. Our analysis reveals that most significant GO-terms we find for the breast cancer network represent functional modules of biological processes that are described by known cancer hallmarks, including translation, immune response, cell cycle, organelle fission, mitosis, cell adhesion, RNA processing, RNA splicing and response to wounding. Furthermore, by using a curated list of census cancer genes, we find enrichment in these functional modules. Finally, we study cooperative effects of chromosomes based on information of interacting genes in the breast cancer network. We find that chromosome 21 is most coactive with other chromosomes. To our knowledge this is the first study investigating the genome-scale breast cancer network.

9. Our Proposal

The idea behind our work is to form a database for lung cancer with information gathered from open information and use that information to correlate the datasets so that we can use a weighted gene regulatory network analysis method to find the module and from the modules the genes responsible for cancer formation.

10. Details

The datasets gathered from multiple open sources was first converted to csv (comma separated value) format. Then the datasets were cross-matched using the R language.

To merge the datasets and do cross-matching among them so that information that can be correlated can be filtered out from it. Afterwards the filtered information will be used to correlate the datasets.

The correlation will be done using a weighted gene co-expression network analysis method that has been redesigned to fit the dataset created. From WGCNA the necessary modules will be separated from which finally the genes responsible will be found.

11. Experiment

The data gathered has been in tab delimited value format, excel sheet format comma delimited value format and there were a bunch other formats as well all those data had to be converted into a comma separated value format which helped the access of data using R. Afterwards some library functions started to malfunction. The root of the cause was identified as a system error in the WGCNA package for R which contained the “Go.db” and the “org.Hs.eg.db” databases. The version of R used in the time of experimentation was 3.5.1 for which apparently the WGCNA package was not updated yet. Therefore, core library functions had to be built again. Afterwards when the formatting of the data was done and the library rebuilt as well, the correlation began. To do the correlation the weighted gene correlation method had been redesigned to perform operation on the dataset designed. There were layered result as the code had been done in the similar fashion the results were analyzed to identify the necessary modules and from the modules the culprit modules which gave the culprit genes as the vastly connected genes in the identified module.

12. Results

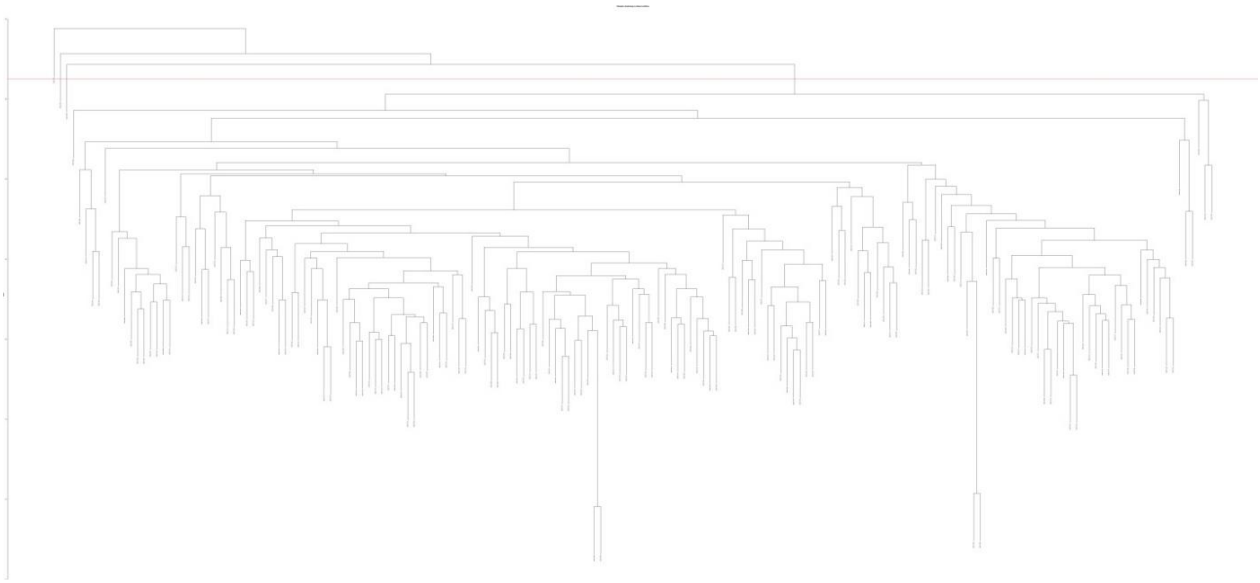

Figure: Hierarchical clustering of gene-expression data of cancer affected patients

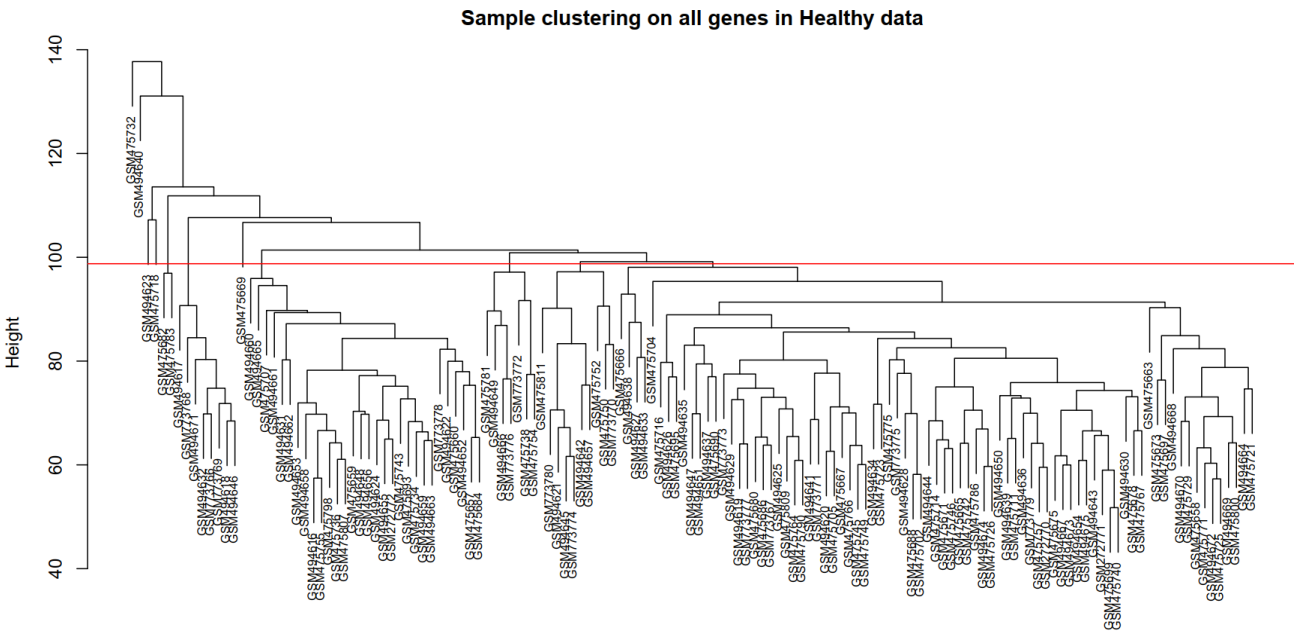

Figure: Hierarchical clustering of gene-expression data of healthy patients

The hierarchical clustering had been used to identify if any pruning of genes had to be done and if so where should it be. All genes under the projected red line had been used in this case to perform operation to correlate.

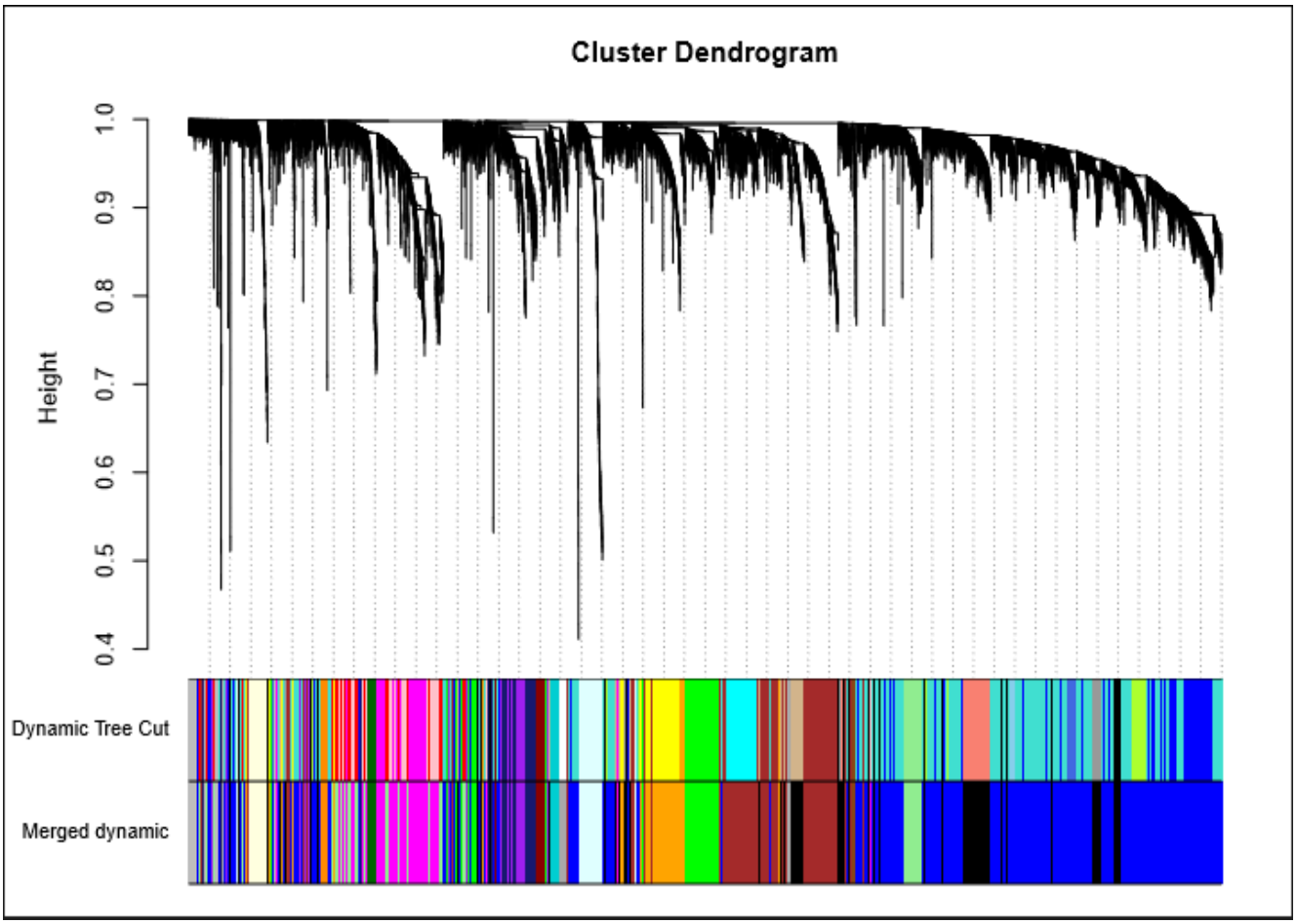

Figure: The hierarchical cluster-tree

After pruning a new tree is generated, divided the clusters and separated them by creating modules and assigning different colors to it. Afterwards these modules are compared with the trait information available as the trait data for patients.

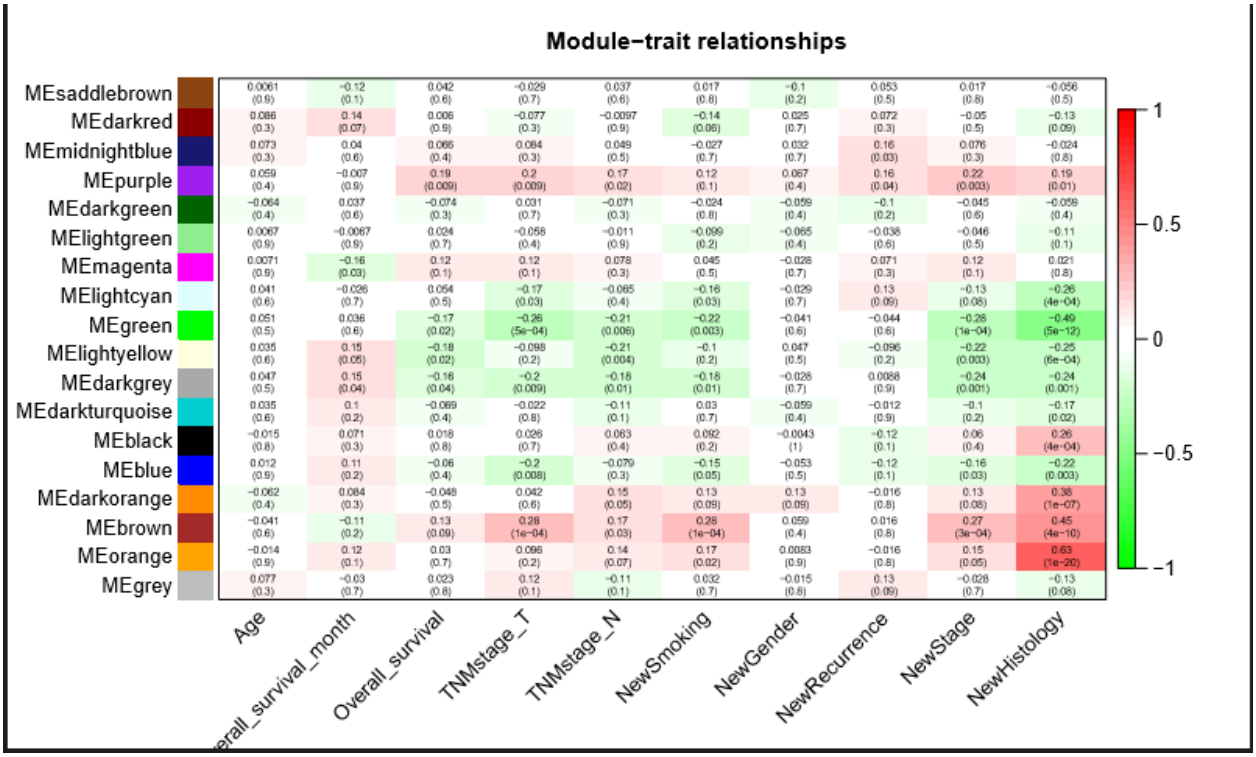

Figure: Module-Trait relationships

The module-trait relationship shows which module has what kind of concentration at different traits. Traits are mapped for both cancer-affected patients and healthy patients. Later the module-trait relation of the cancer-affected patients are matched with the module-trait relation of the healthy patients.

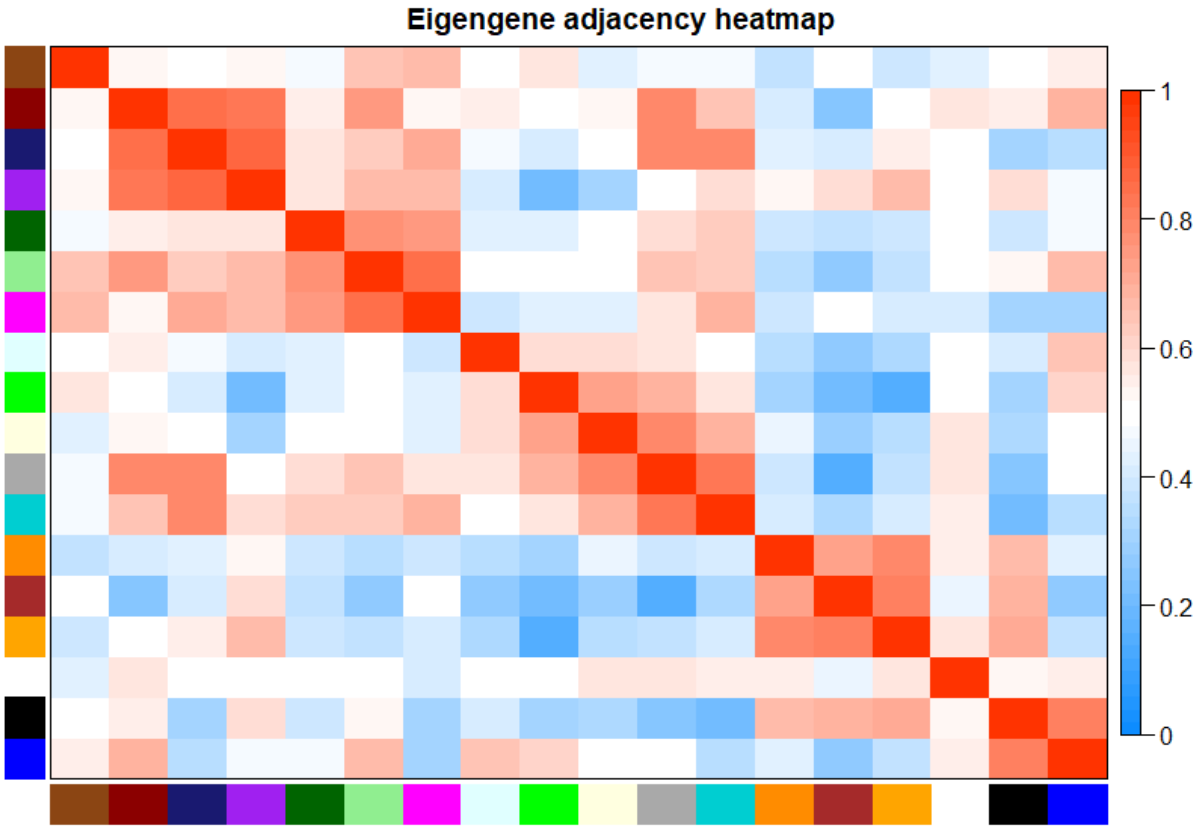

Figure: The heatmap made using the module-trait relationship of the cancer affected patients

The heatmap correlates between the modules and show an inter-modular relationship between them which helps to build a gene regulatory network of the concerned genes. Here the concentration along the diagonal is high due to the modules being absolutely the same. Which is why the upper triangular matrix is where we put our focus in.

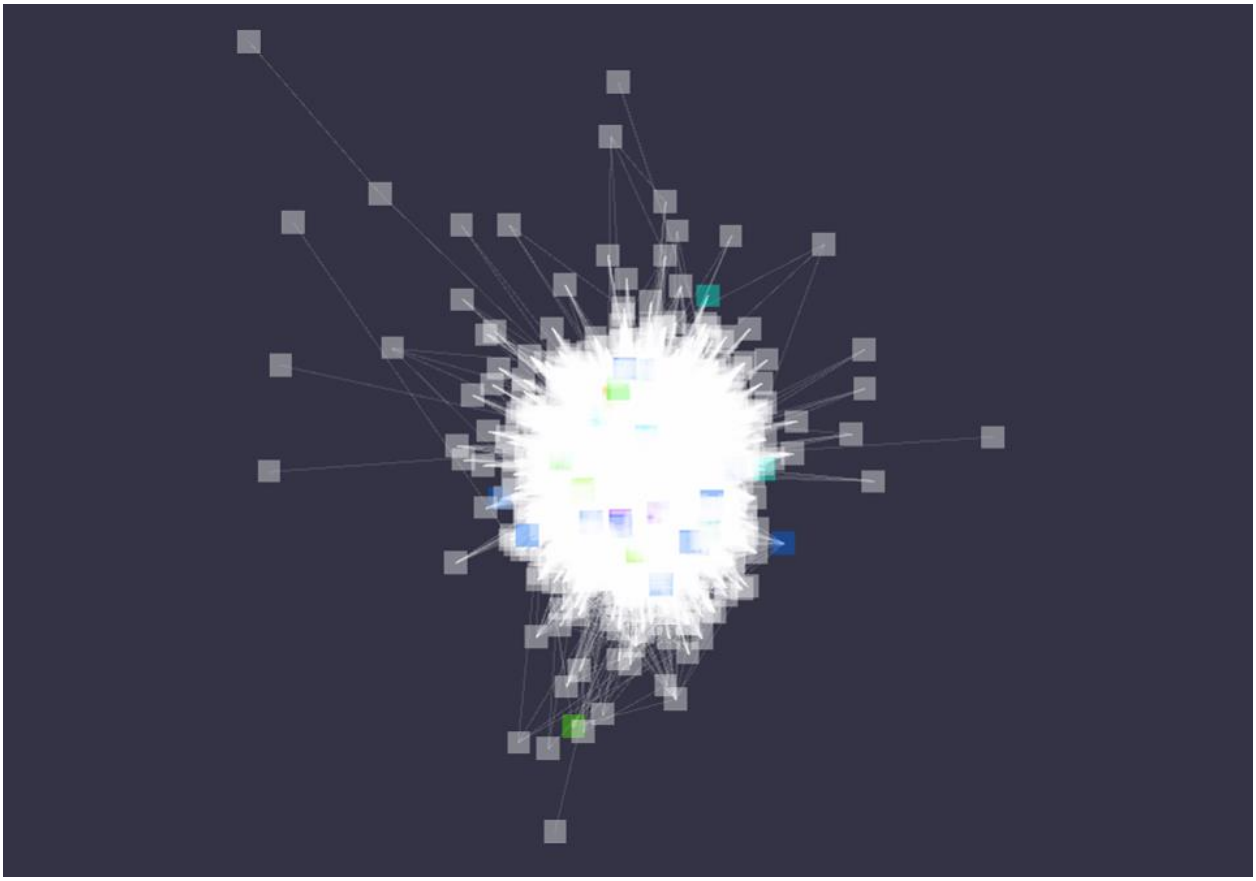

Figure: The gene regulatory network formed using the data from cancer affected patients.

There is a similar set of decisional results developed for the healthy patients as well. The correlation among them has been done using the module-trait relation between the two kinds of datasets. The reason being to identify cancer affected patients by their traits. So that, in the future lung cancer patients can be detected using their traits.

A consensus module has been built that correlates the data between the module-trait relationship between the healthy and cancer affected patients. Afterwards the cancer-affected patient's module-trait data has been cross-matched with the data consensus module-trait data to find the most affected region.

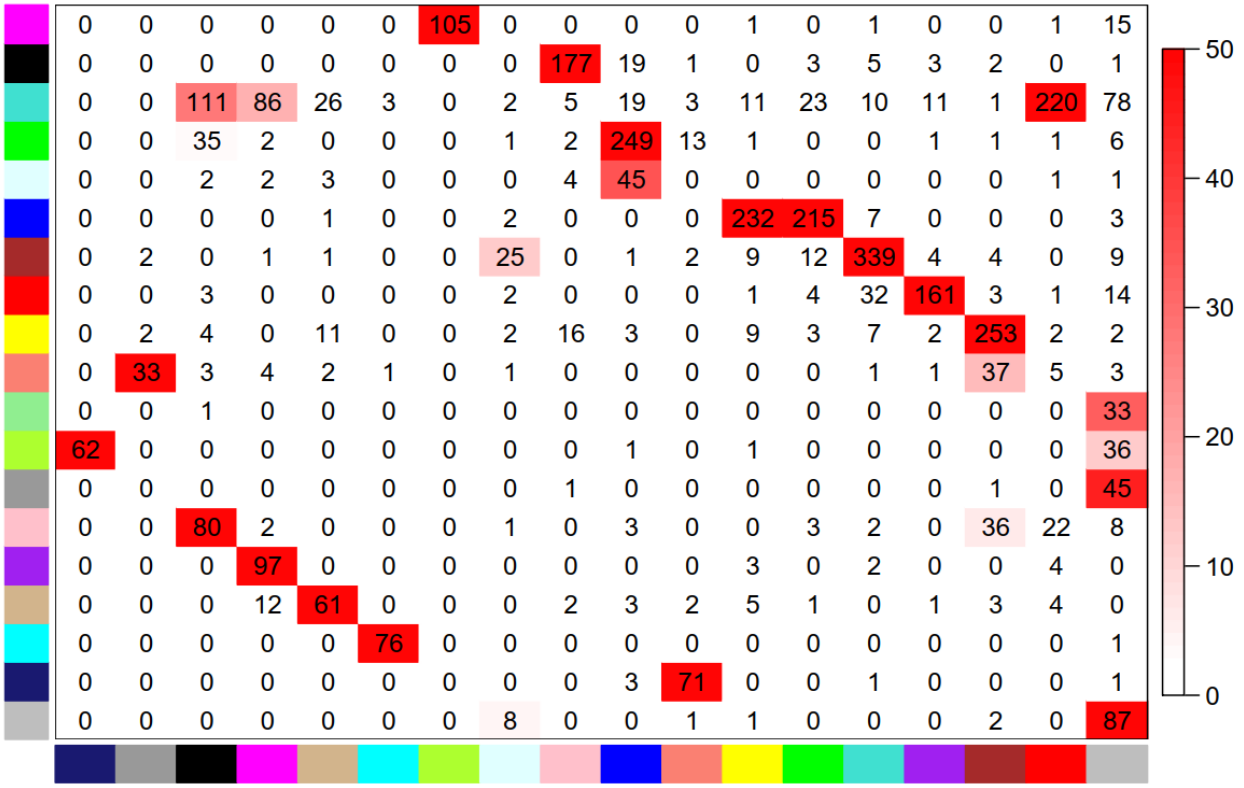

Figure: Relation between module-trait relationship of cancer-affected patients and consensus module

From the above module-trait relationship diagram of consensus module relationship between cancer affected patients to healthy patients to cancer affected patients

13. Result Analysis

The relationship of cancer affected patient to the consensus module trait shows the module that contains the genes that are responsible for the formation of lung cancer. Later a gene regulatory network is formed using the weighted gene co-expression network analysis method to find out the most connected or hub genes. These genes are identified as the culprit genes responsible for the cancer formation. There are some previously identified genes responsible for lung cancer. The genes found from this approach are later compared with the genes that are previously believed to be true. There are some similarities between the two sets. Dissimilarities exist too. However, the similarity between the proven to be guilty genes and the genes identified by this approach suggests that this approach works. Although the percentage of efficiency is yet to be calculated as there exists no benchmark list that claims to have identified all the genes responsible for lung cancer.

14. Conclusion

We have developed a tool that aims to identify unique sub-networks and genes based upon a number of related studies. We explore networks and genes that are robust and unique to a pre-selected number of studies. We support our results using prediction accuracy and a score to test the significance of identifying a subset of unique genes. Furthermore, we created an application interface which allows the user to combine different studies. Based on the findings we conclude that our research is a robust and reliable method to analyze sets of data from lung cancer. It detects the harmful genes responsible for lung cancer that could be potential targets for further research.

15. References

1. Yoli Shavitab, Boyan Yordanovb, Sara-Jane Dunnb, Christoph M. Wintersteigerb, Tomoki Otania, Youssef Hamadib, Frederick J. Liveseya, Hillel Kuglerbc, Automated Synthesis and Analysis of Switching Gene Regulatory Networks, Biosystems 146(2016)26-34.

2. S.-J. Dun et al, Defining an essential transcription factor program for naïve pluripotency, Science 1156 (2014), 10.1126/science.1248882.

3. Boyan Yordanov, Sara-Jane Dunn, Hillel Kugler, Austin Smith, Graziano Martello and Stephen Emmott, a method to identify and analyze biological programs through automated reasoning, 16010; 10.1038/npjsba.2016.10.

4. Yoli Shavit, Boyan Yordanov, Sara-Jane Dunn, Christoph M. Wintersteiger, Youssef Hamadi, and Hillel Kugler, University of Cambridge, UK Microsoft Research, Bar-Ilan University, Israel, Switching Gene Regulatory Networks, 10.1007/978-3-319-23108-2_11.

5. SOMKID INTEP, DESMOND J. HIGHAM, XUERONG MAO, SWITCHING AND DIFFUSION MODELS FOR GENE REGULATION NETWORKS, 10.1137/080735412.