Finding Correlations Of Common Gene Expressions Of Multiple Genetic Diseases By Using Microarray Data

Şaban Dalaman
School of Natural Sciences and Engineering
Istanbul Şehir University
Istanbul, Turkey
Email: sabandalaman@std.sehir.edu.tr

Ali Çakmak
School of Natural Sciences and Engineering
Istanbul Şehir University
Istanbul, Turkey
Email: alicakmak@sehir.edu.tr

Abstract—The gene regulatory network (GRN) reveals the regulatory relationships among genes and can provide a systematic understanding of molecular mechanisms underlying biological processes. The idea studied here is to compare traditional analysis of microarray gene expression profiles and network analysis of GRN inferred using gene expression data. The target is to find, if there is, common gene patterns effective genetic diseases. In this work we applied statistical analysis and clustering methods like k-means and random forest to find common structures among disease gene profiles. The same data is used to create a GRN network and network theory methods applied to find network statistics to inspect underlying mechanism of gene regulatory network for genetic diseases.

I. INTRODUCTION

The ultimate goal of the genomic revolution is to understand the genetic causes behind phenotypic characteristics of organisms. To understand means having a blueprint that opens the exact ways in which genetic components, like genes and proteins, interact to make a complex living system. The availability of genome-wide gene expression technologies has opened ways more easily to identify the interactions between genes in a living system, or gene networks. Gene networks can be modelled and simulated using various approaches. Once the model has been chosen, the parameters need to be fit to the data. Even the simplest network models are complex systems involving many parameters, and fitting them is a non-trivial process, known as *network inference*, *network identification*, or *reverse engineering.*

In this paper we approached the problem of identifying gene relations and patterns common in genetic diseases. We have chosen two ways to deal with this problem. One is traditional statistical analysis and applying clustering methods. The other one is rather new and different method. It is by using network theory to analyze gene regulatory relations by using microarray data.. We start first giving some preliminary definitions.

A. Gene Regulatory Network

Gene regulatory network (GRN) is a network that describes how genes regulate each other by activation or inhibiton through different transcription factors. Transcription factors are proteins that binds to a specific DNA sequences to activate or supress the production of RNA polymerase. GRN is a matrix of links with weights by which all genes are connected each other. Genes are nodes and the links between them are edges in GRN. GRN data can be from intervention data known as knock-down and knock-out data or observational gene expression data as steady-state or time-series. Different approaches can be used to infer GRN such as correlation and information theoretic approches, Bayesian Networks, Dynamic Bayesian Network, Non-Traditional models like ANN,Fuzzy Systems and Evolutionary algorithms or swarm intelligence optimizations (PSO, ACO)

B. DNA Sequencing and Microarray Technology

DNA sequencing is a tool to determine the precise sequence of DNA bases in a sample. Sequencing can yield a wealth of information concerning gene architecture, the control of gene expression, as well as protein structure. Most genes are present in the same quantity in every cell. However, the level at which a gene is expressed, as indicated by mRNA quantities, can vary widely. Gene-expression patterns vary from cell type to cell type. Even within the same cell, gene expression levels may vary as the cell responds to changes in physiological circumstances. One of the most powerful methods for this purpose developed to date is based on DNA hybridization and is called the *DNA microarray* or *gene chip.* This technique enables to monitor the expression levels of a large number of genes simultaneously and provides a global view of gene expression information of the organism under study. Depending upon the specific technology used, DNA microarrays can reflect either absolute expression levels (e.g., Affymetrix GeneChip® arrays) or relative expression ratios (e.g.cDNA microarrays).

II. Related Works

In this section we overview some of the works to show how microarray data has been used to inspect different aspects of gene sequences and patterns like important gene groups. The methods used are from wide range of algorithms like clustering with k-means or hierarchical clustering, classification by using SVM with different kernels, and regression analysis as well as specific statistical tools.

There are many machine learning techniques that can be used in microarray data analysis. Mainly they can be divided as supervised and unsupervised methods. One of them is cluster analysis. David J. Hand and Nicholas A. Heard [1] in their paper have reviewed clustering and pattern discovery methods to find groups and subgroups in microarray data. It seems crucial and fundamental to group or partition objects for human understanding. The most commonly used partitioning method is "k-means" clustering. Cluster analysis with gene expression data has its own aspects, such as high dimensionality and low number of cases for some problems. Variable selection is a critical problem while classifying cells or samples, on the basis of the genes (and hence requires a very large number of variables) but unimportant when one is trying to classify genes themselves, perhaps on the basis of very few expression conditions.

As an examples of classifying, this paper [2] presents methods for analyzing gene expression data to classify cancer types. The techniques, such as Bayesian networks, neural trees, and radial basis function (RBF) networks, are used for the analysis of the CAMDA Data Set 2. The Bayesian network represents the joint probability distribution for a set of random variables efficiently based on the concept of conditional independence. Once the Bayesian network whose nodes represent gene expression levels and the cancer class label is constructed from the gene expression data, the probability of the cancer class label given some gene expression levels for a new sample can be inferred. Neural tree models represent multilayer feedforward neural networks as tree structures. Among these machine learning techniques, Bayesian network learning has the power to capture the relationships among genes in comprehensible format. Neural tree learning seems the best in finding out a small set of interesting genes for effective classification.

In this work [3] Markov cluster algorithm (MCL), Molecular complex detection (MCODE) and Clique percolation method (CPM) were used to decompose human PPI network into dense clusters as the candidates of disease-related clusters.

SVM is widely used method in studying patterns of gene regulatory network. [5] is one of them. SVMs are particularly well to the analysis of broad patterns of gene expression from DNA micro-array data because they can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients). They integrate pattern selection and feature selection in a single consistent framework.

Network-based approach recently attracted much attention in gene expression analysis. [4] is an examples of these works. In this study it was used a network-oriented and data-driven bioinformatic approach that searches for association of genes and diseases based on the analysis of genome-wide expression data derived from microarrays or RNA-Seq studies. This bioinformatic approach was implemented in an R package, named geNetClassifier, available as an open access tool in Bioconductor.

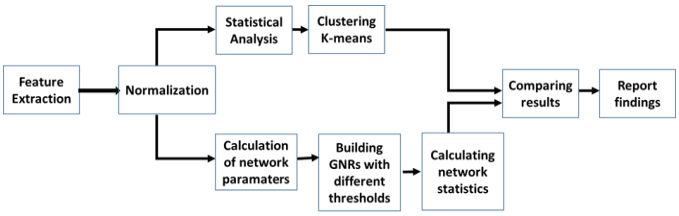
In the paper [6] the authors present a network-based framework to identify the location of disease modules within the interactome and use the overlap between the modules to predict disease-disease relationships. Using the methods of network science and machine learning, they show that disease modules for diseases whose number of associated genes exceeds a critical threshold determined by the network incompleteness can be uncovered. The findings show that disease proteins associated with 226 diseases are clustered in the same network neighborhood, displaying a statistically significant tendency to formidentifiable disease modules. [7],[8],[9],[10],[11],[12] are also examples of network-based approachs to show its power in the analysis of gene expression data.

III. Proposed Method

In most of the related works, either GRN from samples of human tissues for a disease or several GRNs for each disease from human tissues are inferred. Then comparing the analysis of these inferred networks , gene expressions are investigated.

In this work, the aim was two-fold. First the data sets were built for genetic diseases by using microarray samples of human tissues. At the same time a combined GRN was inferred using the same microarray samples. The constructed GRN was considered as hiding the valuable information about the mutual-effects of genes for the all diseases mentioned. Then first the data sets were analysed by using k-means clustering and random forest algorithms to find clusters of genes and gene-relation active for these diseases at the same time. Second GRNs constructed with these samples and correlation between gene expressions were analyzed by network analysis tools to get important network indicators. The clustering findings were compared with the interpretations of network analysis results.

The main steps of the proposed algorithm are outlined as follows.



A. Preparation of datasets

Several human disease tissues samples were downloaded from NCBI. The disease names are **asthma, leukaemia, diabetes, Parkinson, pancreatic cancer, lung cancer and breast cancer**. The data sets was pre-processed by filling missing values and normalized. All the datasets were from the same kind of microarray platform so that comparing them would be easy because they have the same gene annotations.

B. Analysis of datasets

Datasets were analysed by using clustering algorithm k-means and random forest algorithms. Each gene expression level is a vector comprised of values from diseases sample data. Then similarity values and clustering values were calculated. The purpose was here to find the most significant gene groups that are active for all diseases and their role on the regulation of other genes.

C. *Building the combined disease GRN network*

Pearson correlation coefficient (PCC) scores between genes called as co-expression similarity were calculated by using expression levels in different samples datasets. Using a threshold procedure, the co-expression similarity value was transformed into a measure of connection strength. Then PCC scores of gene pairs were used to construct symmetric and undirected network with no self-connections. The correlation is +1 if there is a perfect positive linear relationship, −1 if there is a perfect negative linear relationship and values between −1 and 1 indicates the degree of linear dependence between the variables. Closer the coefficient to either −1 or +1, stronger the correlation between the variables. If the coefficient is zero, the variables are independent. Once, the pair-wise correlation coefficient between genes were computed, next those coefficients having absolute values above a threshold were selected and eliminated weakly correlated gene pairs. With different threshold values, GRN edge weights were adjusted and different GRNs were built. The method of changing threshold value was used to find robust network structure forming the final GRN for topological analysis.

D. Visualization of GRNs

By using network visualization tools, GRNs built for each sample and clustering diagrams were shown and compared.

E. Calculating network statistics and Topologic analysis

GRN was considered as a social network showing deep connections and relations between gene-pairs and gene clusters. Geometric interpretation of GRN analysis was based on the assumption that wide-range of network statistics listed below can be used to find characteristics of gene co-expression relations thus showing underlying connections between diseases.

The following network statistics were examples of key indicators of gene co-expression network to be calculated:

Node connectivity showing gene activity level, maximum adjacency ration to determine where a hub gene is formed , density to find a subnet of genes , network centralization to describe properties of clusters, network heterogeneity for a degree of scale-free topology, clustering coefficient that means the higher the clustering coefficient of an individual, the higher the affection among genes.

Employing networks statistics and comparing them with clustering statistics, important gene or gene clusters were to be identified. These results may show gene clustering and their similarities among these diseases.

IV. EXPERIMENTAL WORK

The disease names were selected as **asthma, leukaemia, diabetes, Parkinson, pancreatic cancer, lung cancer and breast cancer**. 10 human genetic disease tissue samples were downloaded from NCBI GEO Database. There were total 70 samples for 22646 genes. All the datasets are from the same kind of microarray platform so that comparing them would be easy because they have the same gene annotations. Annotation platform is GPL97. Annotation platform title is [HG-U133B] Affymetrix Human Genome U133B Array and annotation platform organism is Homo sapiens.

Accession codes for samples as follows:

Diabetes Samples: GSM254184, GSM254185, GSM254187, GSM254189, GSM254190, GSM254186, GSM254188, GSM254194, GSM254195, GSM254196

Leukaemia samples: GSM559433, GSM559434, GSM559436, GSM559437, GSM559438, GSM559440, GSM559441, GSM559442, GSM559444, GSM559445

Parkinson Samples: GSM208700, GSM208701, GSM208702, GSM208703, GSM208704, GSM208705, GSM208706, GSM208707, GSM208708, GSM208709

Asthma Samples: GSM3922, GSM3924, GSM3926, GSM3928, GSM3930, GSM3932, GSM3934, GSM3936, GSM3938, GSM3940

Pancreatic Cancer Samples: GSM1060008, GSM1060009, GSM1060010, GSM1060011, GSM1060019, GSM1060020, GSM1060021, GSM1060022, GSM1060023, GSM1060024

Lung Cancer Samples: GSM783029, GSM783030, GSM783031, GSM783032, GSM783033, GSM783034, GSM783035, GSM783036, GSM783037, GSM783038

Breast Cancer Samples: GSM151910, GSM151911, GSM151912, GSM151913, GSM151914, GSM151915, GSM151916, GSM151917, GSM151918, GSM151919

Using these samples, 8 data files were created. One for each disease and one for combination of them. In the data sets rows were genes and columns from samples accession ids. Gene names were formed by combining gene symbol name and microarray probe id to get unique names. Gene expression values from microarray data were normalized by first applying log normalization and then z-score conversion for each gene values. This procedure was applied for each data set separately.

A. Gene Rank Ordering

Gene rank orders were found by first calculating variance of samples values for each gene and ordering them.

Figure

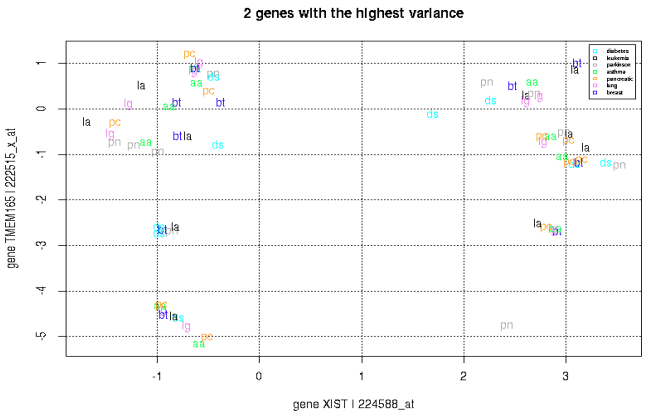


Fig 1. This diagram for two genes with highest variance.

The following lists show the list of 10 top genes in the rank order.

	Disease "ALL"	"Diabetes"
Gene 1	"XIST | 224588_at"	"XIST | 224588_at"
Gene 2	"TMEM165 | 222515_x_at"	"XIST | 224590_at"
Gene 3	"SYCN | 229995_at"	"XIST | 227671_at"
Gene 4	"AFFX-r2-Bs-dap-M_at | AFFX-r2-Bs-dap-M_at"	"XIST | 224589_at"
Gene 5	"MIR675///H19 | 224997_x_at"	"231597_x_at | 231597_x_at"
Gene 6	"AFFX-TrpnX-3_at | AFFX-TrpnX-3_at"	"LOC100996809///HLA-DRB5///HLA-DRB4///HLA-DRB3///HLA-DRB1 | 238900_at"
Gene 7	"SOX2-OT | 231898_x_at"	"LOC100509457///HLA-DQA1 | 236203_at"
Gene 8	"AFFX-r2-Bs-dap-3_at | AFFX-r2-Bs-dap-3_at"	"TXLNGY | 223645_s_at"
Gene 9	"MS4A1 | 228592_at"	"233847_x_at | 233847_x_at"
Gene 10	"AFFX-TrpnX-M_at | AFFX-TrpnX-M_at"	"241868_at | 241868_at"

	Disease "Leukaemia"	"Parkinson"
Gene 1	"CTLA4 | 236341_at"	"XIST | 224588_at"
Gene 2	"MS4A6A | 223280_x_at"	"XIST | 227671_at"
Gene 3	"MS4A6A | 224356_x_at"	"HIPK2 | 225116_at"
Gene 4	"LOC100509457///HLA-DQA1 | 236203_at"	"USP31 | 226035_at"
Gene 5	"240666_at | 240666_at"	"PWAR6 | 226587_at"
Gene 6	"238712_at | 238712_at"	"TOB1 | 228834_at"
Gene 7	"GNB4 | 225710_at"	"C7orf61 | 229913_at"
Gene 8	"SNORD3D///SNORD3C///SNORD3B-2 ///SNORD3A///SNORD3B-1 | 235102_x_at"	"MIR612///NEAT1 | 227062_at"
Gene 9	"RBM20 | 238763_at"	"ZNF275 | 225383_at"
Gene 10	"GPAT2 | 235557_at"	"NOVA2 | 235560_at"

	Disease "Asthma"	"Pancreatic Cancer"
Gene 1	"XIST | 224588_at"	"INHBA | 227140_at"
Gene 2	"ACTB | 224594_x_at"	"XIST | 224588_at"
Gene 3	"XIST | 224589_at"	"SYCN | 229995_at"
Gene 4	"XIST | 224590_at"	"SERPINB6 | 231628_s_at"
Gene 5	"239591_at | 239591_at"	"COL8A1 | 226237_at"
Gene 6	"AFFX-hum_alu_at | AFFX-hum_alu_at"	"CTHRC1 | 225681_at"
Gene 7	"LOC100996809///HLA-DRB5///HLA-DRB4///HLA-DRB3 ///HLA-DRB1 | 238900_at"	"TMED6 | 236430_at"
Gene 8	"ACTB | AFFX-HSAC07/X00351_M_at"	"ANTXR1 | 224694_at"
Gene 9	"SNORD24///SNORD36A///SNORD36B///RPL7A | 224930_x_at"	"COL3A1 | 232458_at"
Gene 10	"241647_x_at | 241647_x_at"	"ERP27 | 227450_at"

	Disease "Lung Cancer"	"Breast Cancer"
Gene 1	"INHBA | 227140_at"	"XIST | 224588_at"
Gene 2	"XIST | 224588_at"	"AFFX-HUMRGE/M10098_5_at | AFFX-HUMRGE/M10098_5_at"
Gene 3	"SYCN | 229995_at"	"EHF | 225645_at"
Gene 4	"SERPINB6 | 231628_s_at"	"CYBRD1 | 222453_at"
Gene 5	"COL8A1 | 226237_at"	"NARR///RAB34 | 224710_at"
Gene 6	"CTHRC1 | 225681_at"	"AFFX-r2-Hs18SrRNA-5_at | AFFX-r2-Hs18SrRNA-5_at"
Gene 7	"TMED6 | 236430_at"	"AFFX-r2-Hs18SrRNA-M_x_at | AFFX-r2-Hs18SrRNA-M_x_at"
Gene 8	"ANTXR1 | 224694_at"	"CCDC88A | 225045_at"
Gene 9	"COL3A1 | 232458_at"	"CAPS | 231729_s_at"
Gene 10	"ERP27 | 227450_at"	"XIST | 224590_at"

B.K-Means Clustering

The top 500 gene were selected from the list ordered by variance. The gene lists were used in k-means clustering. The best choice for the number of clusters were found by using total within-cluster sum of squares.

According to this metric, the best number of clusters for all samples and combined samples is 3.

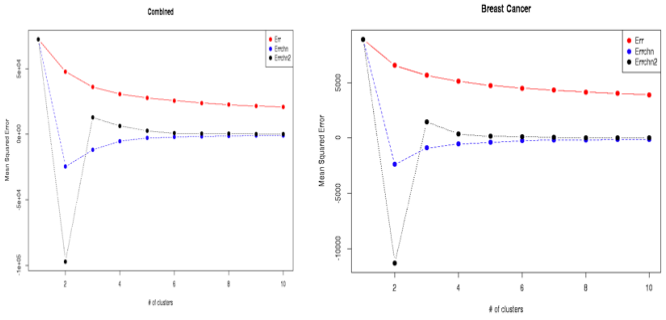


Fig 2. Example plots for total within-cluster sum of squares.

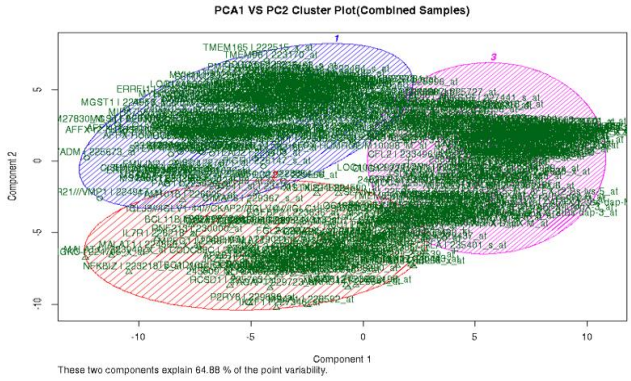


Fig 3. PCA1 and PCA2 plot for k-means clusters

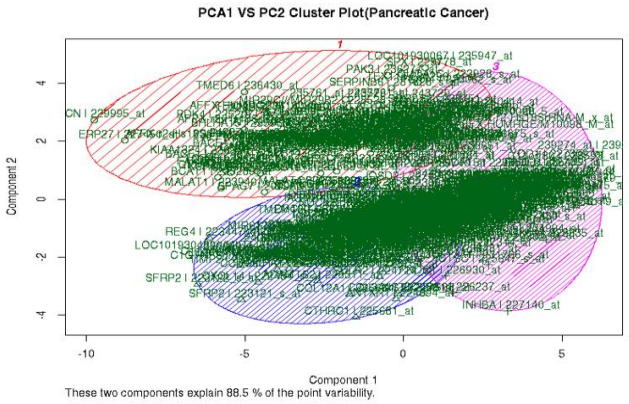


Figure 4. PCA1 and PCA2 plot for k-means clusters

PCA scatter plots show how gene expression values are clustered across PCA values.

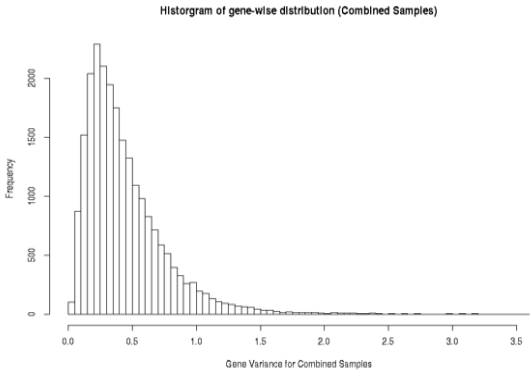


Fig 5: Histogram of gene variances for combined set

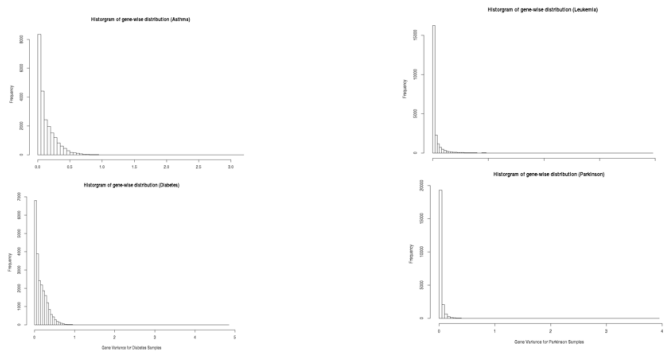


Fig 6: Histogram of gene variances for sample sets

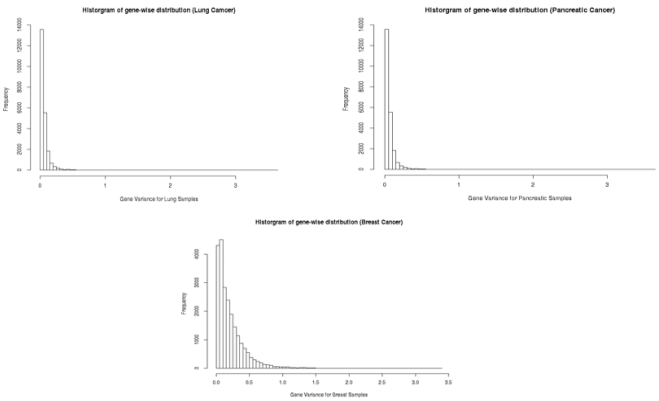


Fig 7: Histogram of gene variances for sample sets

Histogram of gene variances shows only small number genes higher activity as expected.

C. Random Forest Clustering

The top 500 genes from the list ordered by variance were selected and used in random forest clustering. By using Random forest algorithm as an unsupervised method, gene importance values from gini index were calculated for each sample set and combined set. The importance values can be interpreted as showing most active genes for each disease case.

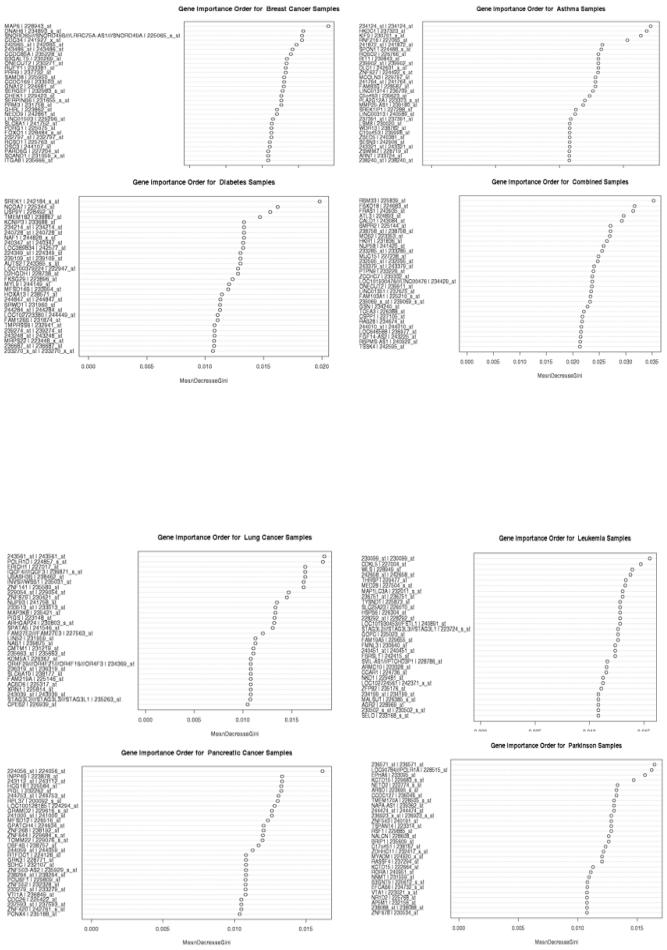


Fig 8: Gene importance list for all sample sets and combined set.

D. Network Analysis

As explained before, GRN nodes and links were found by using Pearson correlation of each gene pairs. Then with threshold value of 0.85, gene-pairs were selected and GRN node and edge lists were formed. By using node and edge lists, an undirected network was created for each sample and combined set.

Fig 9. Network layouts for disease samples

Table1. Network Properties

Properties	ALL	Diabetes	Leukaemia	Parkinson
# of nodes	452	353	420	423
# of links	9660	846	2474	6336
Link density	21.37	2.39	5.89	14.97
Graph connectance	0.047	2.396	5.890	0.035

Properties	Asthma	Pancreatic	Lung	Breast
# of nodes	372	497	382	406
# of links	896	77310	1864	2982
Link density	2.40	155.55	4.87	7.34
Graph connectance	0.006	0.313	0.012	0.018

From the network properties, it can be seen that Parkinson samples show very high link density compared to other samples. However its graph connectance is not much relative to other. This means although there are many active genes but their correlation is not high. Diabetes and Leukaemia samples has higher graph connectance indicating high correlation between genes.

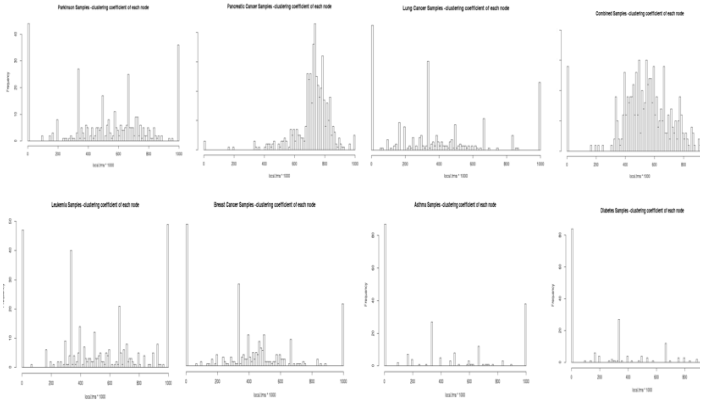
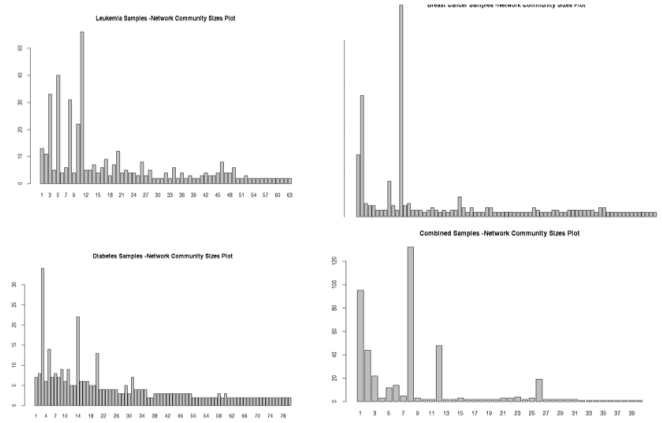


Fig 10. Clustering coefficient for each node



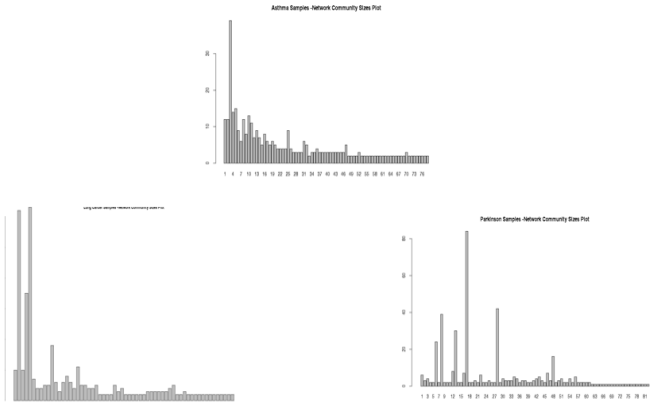


Fig 11. Network community size histograms

Community shows which gene groups are active for each disease case. The community size values show that the number of communities with higher sizes are very low. This means small number of communities with relatively high sizes are active.

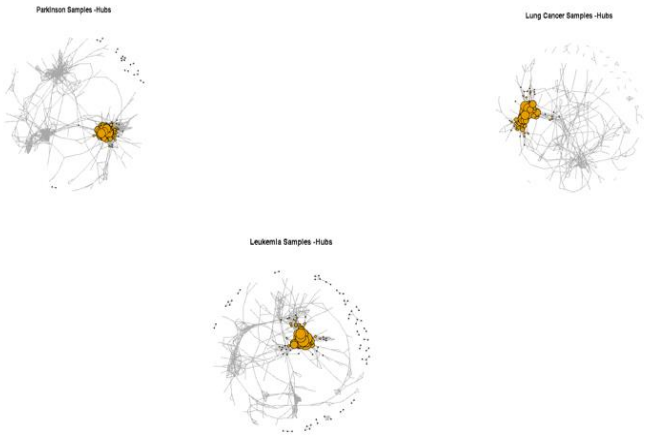


Fig 12. Node hub scores layout for each network

Gene with high hub score has higher number of connections to other nodes. This score shows its activity level for each disease case. These gene list can be compared with random forest clustering result to see their results compatibility.

Table2. Top 5 genes with highest hub scores

Combined Samples

Gene 1	"244544_at | 244544_at"
Gene 2	"NMUR2 | 224088_at"
Gene 3	"CFAP221 | 231043_at"
Gene 4	"DCAF8 | 238338_at"
Gene 5	"231205_at | 231205_at"

Diabetes Samples

Gene 1	"XIST | 224590_at"
Gene 2	"XIST | 224588_at"
Gene 3	"XIST | 227671_at"
Gene 4	"TXLNGY | 223645_s_at"
Gene 5	"TXLNGY | 223646_s_at"

Leukaemia Samples

Gene 1	"236417_at | 236417_at"
Gene 2	"RGCC | 239827_at"
Gene 3	"230014_at | 230014_at"
Gene 4	"239504_at | 239504_at"
Gene 5	"HBP1 | 236645_at"

Parkinson Samples

Gene 1	"ANKIB1 | 224687_at"
Gene 2	"HACD3 | 234000_s_at"
Gene 3	"CLOCK | 225856_at"
Gene 4	"RC3H2 | 231716_at"
Gene 5	"MIR612///NEAT1 | 234989_at"

Asthma Samples

Gene 1	"ACTB | AFFX-HSAC07/X00351_M_at"
Gene 2	"ND4 | 224372_at"
Gene 3	"SNORD24///SNORD36A///SNORD36B///RPL7A | 224930_x_at"
Gene 4	"RPS10 | 200095_x_at"
Gene 5	"RPL19 | 200029_at"

Pancreatic Cancer Samples

Gene 1	"INHBA | 227140_at"
Gene 2	"LUZP6///MTPN | 224656_s_at" "
Gene 3	"SULF2 | 224724_at"
Gene 4 "	"CTHRC1 | 225681_at"
Gene 5	"ANTXR1 | 224694_at"

Lung Cancer Samples

Gene 1	"PLXDC2 | 236297_at"
Gene 2	"COL1A2 | 229218_at"
Gene 3	"GLT8D2 | 227070_at"
Gene 4	"PLXDC2 | 226865_at"
Gene 5	"TMEM119 | 227300_at"

Breast Cancer Samples

Gene 1 "	"ERBB3 | 226213_at"
Gene 2	"EPB41L5 | 229292_at"
Gene 3	"TC2N | 234970_at"
Gene 4	"228440_at | 228440_at"
Gene 5	"MAL2 | 224650_at"

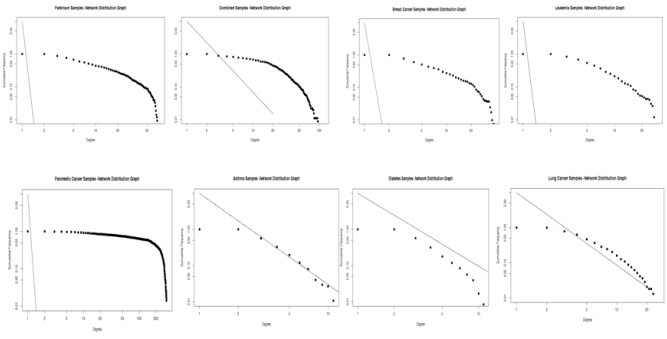


Fig 13: Power law fit for each samples network

Power law fit diagrams show how compatible GRN to the scale-free network phenomenon This relations and network theory methods can be used together to analyse further GRN structures.

VI. CONCLUSION

The purpose of this work is to identify gene relations and their group structures common in genetic diseases. We have selected Diabetes, Leukaemia, Asthma, Parkinson, Pancreatic cancer, Lung cancer and Breast cancer as diseases to study. Microarray data from human tissues with diseases were downloaded NCBI and data sets for each disease and a combination of them. Then by using clustering and network analysis techniques our study was executed. The combined data set was used to search for common structure among these diseases by comparing individual disease samples. The purpose for using clustering and network analysis is to compare their results compatibility and to investigate how they can be used together effectively.

Clustering and statistical analysis are better in finding global structures. However network analysis seems to perform well both local and global analysis. Visualization tools together with network analysis provide better insight for finding common structures. Adding time dimension to network analysis may help to find the evolving structures in GRN. Another important point is domain knowledge. In this study knowing gene functions and their roles in cell machinery is crucial. Microarray data pre-processing and using sophisticated domain specific gene selection methodologies are the most important step Statistical analysis can used for feature selection to find the most relevant gene sets to study. GRN can be inferred by using these sets. In short these methods are not alternatives but complements each other with their powerful aspects.

V. FUTURE WORK

This study may be considered as the beginning phase in establishing a methodology and searching common gene patterns active in genetic diseases. There are important future steps to extend its applicability. One of them is using human normal tissues as control group. Control group clustering analysis can be used to find better feature selection criteria. Another enhancement for feature selection is to use Gene Ontology Database. Domain specific gene selection algorithms, grouping genetic diseases and applying gene selection procedures separately before starting to study the combined sample may improve the applicability of machine learning algorithms more effectively.

[1] Finding Groups in Gene Expression Data, David J. Hand and Nicholas A. Heard, Journal of Biomedicine and Biotechnology. 2005:2 (2005) 215–225 • DOI: 10.1155/JBB.2005.215REFERENCES

[2] Applying machine learning techniques to analysis of gene expression data : Cancer diagnosıs , Kyu-Baek Hwang, Dong-Yeon Cho, Sang-Wook Park, Sung-Dong Kim, and Byoung-Tak Zhang, Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea

[3] Prediction of Human Disease-Related Gene Clusters by Clustering Analysis, Peng Gang Sun, Lin Gao and Shan Han, International Journal of Biological Sciences 2011; 7(1):61-73

[4] Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles, Sara Aibar, Celia Fontanilla, Conrad Droste, Beatriz Roson-Burgo, Francisco J Campos-Laborie, Jesus M Hernandez-Rivas, Javier De Las Rivas, International Conference on the Brazilian Association for Bioinformatics and Computational Biology Belo Horizonte, Brazil. 28-30 October 2014

[5] Gene Selection for Cancer Classification using Support Vector Machines Isabelle Guyon, Jason Weston, Stephen Barnhill, M.D. and Vladimir Vapnik ,Barnhill Bioinformatics, Savannah, Georgia, USA * AT&T Labs, Red Bank, New Jersey, USA

[6] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Gaussian, Marc Vidal, Joseph Loscalzo, Albert-László Barabási, Uncovering disease-disease relationships through the incomplete interactome

[7] Network based analyses of gene expression profile of LCN2 overexpression in esophageal squamous cell carcinoma Bingli Wu, Chunquan Li, Zepeng Du, Qianlan Yao, Jianyi Wu, Li Feng, Pixian Zhang, Shang Li, Liyan Xu & Enmin Li, Scientific Reports 4, Article number: 5403 (2014) doi:10.1038/srep05403

[8] RankGene: identification of diagnostic genes based on expression data, Yang Su T.M. Murali Vladimir Pavlovic Michael Schaffer Simon Kasif, Bioinformatics (2003) 19 (12): 1578-1579.

[9] Thomas J, Olson J, Tapscott S, Zhao L: An efficient an robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Research. 2001, 11: 1227-1236.

[10] Statistical methods and microarray data, Lev Klebanov, Xing Qiu, Stephen Welle & Andrei Yakovlev Nature Biotechnology 25, 25 - 26 (2007) doi:10.1038/nbt0107-25

[11] Finding Groups in Gene Expression Data, David J. Hand* and Nicholas A. Heard J Biomed Biotechnol. 2005; 2005(2): 215–225, doi: 10.1155/JBB.2005.215

[12] Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles, Sara Aibar,1 Celia Fontanilla, Conrad Droste, Beatriz Roson-Burgo, Francisco J Campos-Laborie,1Jesus M Hernandez-Rivas, and Javier De Las Rivas

[13] Practical Graph Mining With R , Chapman & Hall/CRC Data Mining and Knowledge Discovery Series

[14] Gene Expression Studies Using Affymetrix Microarrays, Hinrich Göhlmann Willem Talloen

[15] A Practical Approach To Microarray Data Analysis, Kluwer Academic Publishers

[16] R Graphs Cookbook Second Edition, Packt Publishing

[17] Data Analysis Tools For DNA Microarrays , Sorin Draghici

[18] A Tutorial Review of Microarray Data Analysis, Alex Sánchez and M. Carme Ruíz de Villa

[19] Applied Statistics for Bioinformatics using R, Wim P. Krijnen