GENE CLUSTERING
AND
CONSTRUCTION OF
INTRA-CLUSTER GENE REGULATORY NETWORK

Aparajita Khan

Registration No.: 121256 of 2012-13

Examination Roll No.: M6TCT1518

Department Of Computer Science And Engineering

Jadavpur University

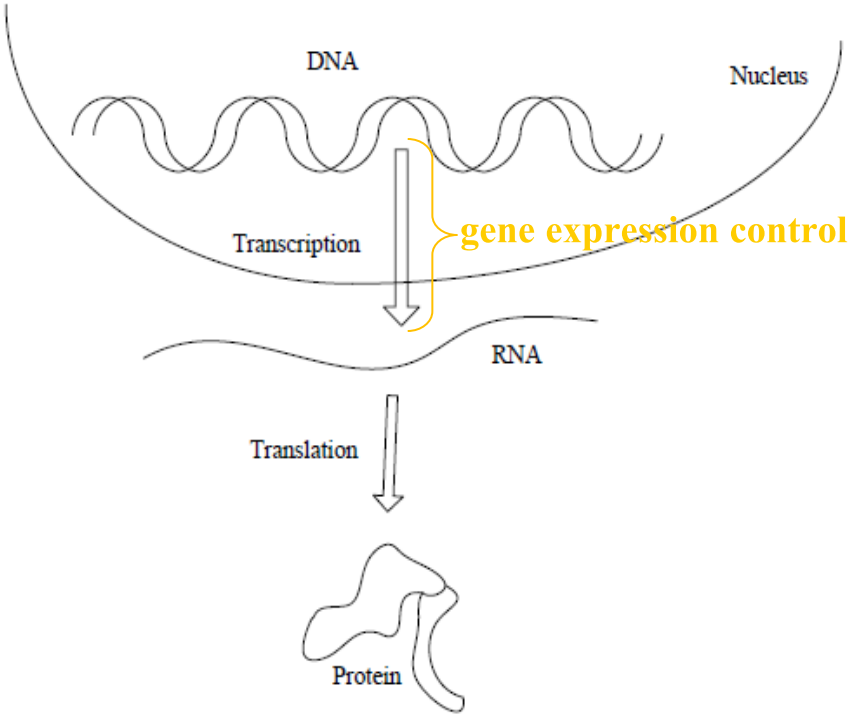
OUTLINE OF PRESENTATION

- Introduction
- Identification of Differentially Expressed Genes
- Clustering of Significant Genes
- Construction of Gene Regulatory Network
- Conclusion
- References

INTRODUCTION

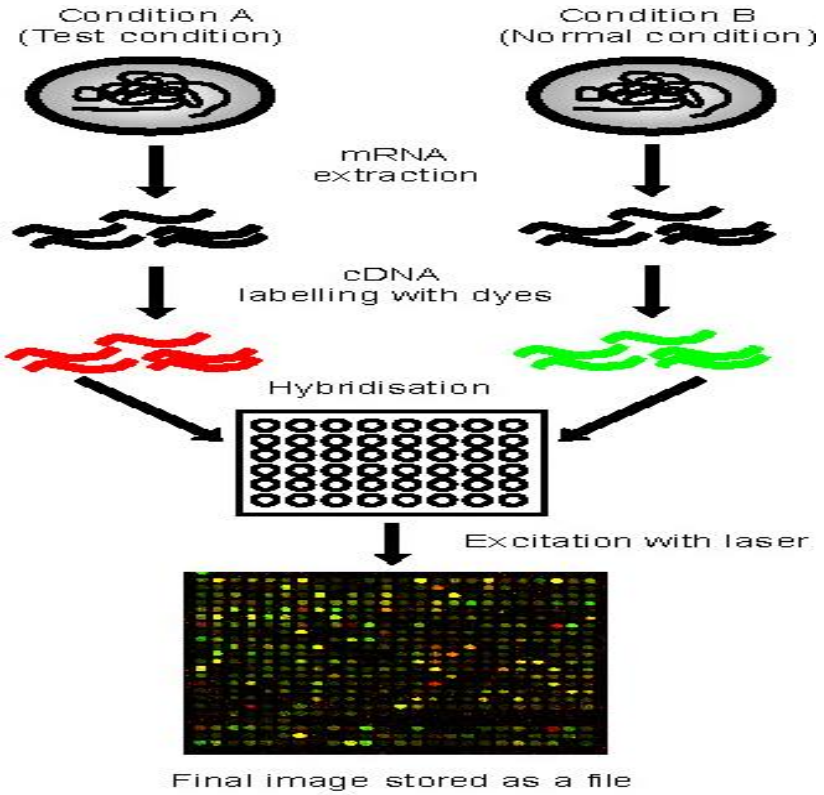
Genes and Gene Expression

- DNA – genetic material that stores and transmits hereditary information from generation to genera
- Gene – segment of DNA that codes for a specific protein
- All cells in an organism have identical genes but gene expression and regulation is responsible for robust dynamic behavior of cell



Microarray Technology

- Monitors genome wide expression profile of organism under study
- Microarray – glass slide with DNA molecules orderly fixed at specific spots
- Spots contain few million copies of DNA corresponding to particular gene
- Works by DNA Hybridization



Motivation of work

- Lung cancer - uncontrolled growth of abnormal cells in one or both lungs.
- Cancer develops due to DNA damage and disruption of control of gene expression and regulation.
- Gene expression analysis provides insight into gene function and disease physiology.

In this work 3 key issues are addressed

- identification of genes exhibiting differential gene expression patterns between diseased and non-diseased population
- clustering significant genes into groups of co-expressed genes
- construction intra-cluster gene- regulatory network

Dataset Under Study

- The study [1] explores patterns of pathway deregulation in normal airway epithelial cells from patients with and without lung cancer.
- Gene Expression Omnibus (GEO) Dataset GDS2771 [2] reports genome wide expression values for 192 smokers with suspect of lung cancer.

2 groups of population in dataset

- Healthy smokers (90 samples)
- Diseased smokers (97 samples)

Gene Subset Selection

- COSMIC, the Catalogue Of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk) manually curates 2002811 coding point mutations in human genes from scientific literature.
- COSMIC's *Cancer Browser* tool reports 24283 genes including alias to be mutated in lung cancer.
- Gene expression profile corresponding to 11237 genes were found in the dataset GDS2771 were considered for further study.

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

Objective

- Differentially expressed genes - gene data determined to be statistical outliers from some standard state, which cannot be ascribed to chance or natural variability.
- Identify genes exhibiting differential expression patterns amongst the two classes of population
- These genes serve as potential pharmaceutical targets and diagnostic markers.

Methodology

- Statistical Hypothesis Testing
- **Null Hypothesis** H_0 : expression levels in healthy and diseased population comes from normal distributions with equal means.
- **Alternative hypothesis** H_A : data comes from different distributions with unequal means.
- For each gene compute Welch's *t*-statistic $t = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{\frac{s_A^2}{N_A} + \frac{s_B^2}{N_B}}}$ where $\overline{X_A}, \overline{X_B}$ are the sample means, and sample variances are s_A^2, s_B^2
- p-value quantifies the probability of observing a test statistic of this extreme or more given that both samples come from the same distribution.
- A small p-value indicates that there is a small chance of getting this data if no real difference existed.

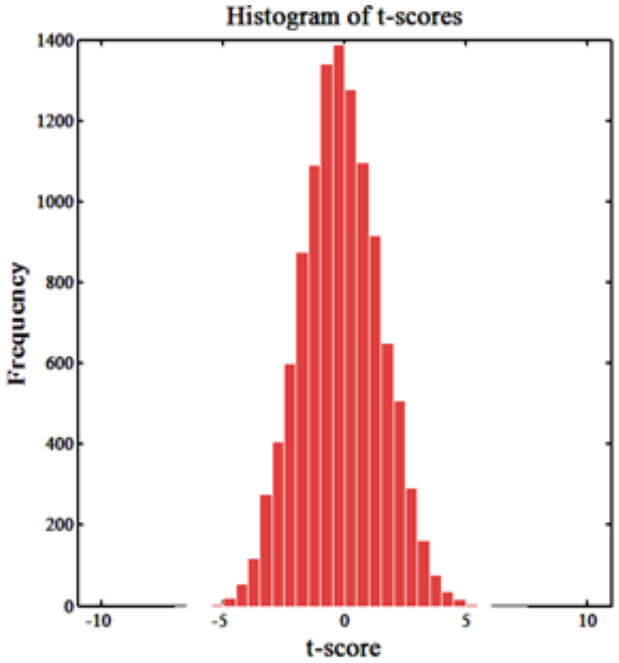
False Discovery Rate (FDR) and q-Value

- p-value measures significance in terms of **false positive rate**.
- p-value cut-off of 0.05 means that there is 5% chance that this gene is called significant when it is truly null. Out of 15000 genes 750 are false positives.
- q-values measures significance in terms of false discovery rate.

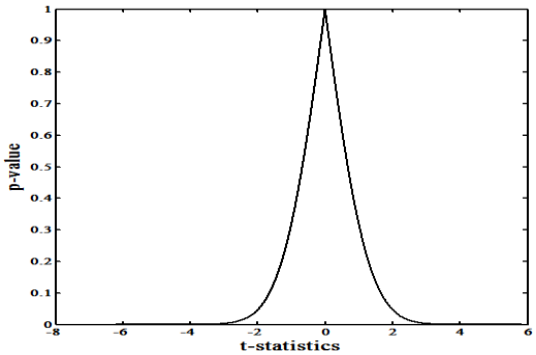
$$FDR = \mathrm{E}\left[\frac{F}{T+F}\right] = \mathrm{E}\left[\frac{F}{S}\right]$$

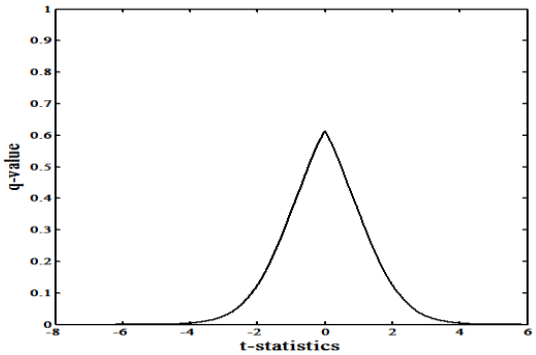
- Takes into account that several features are simultaneously tested, hence a better measure.
- A q-value cut off of 0.05 results in a FDR of 5% among all significant genes.

Graphical Plots

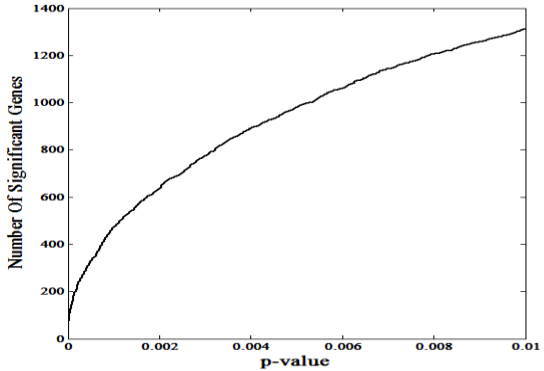


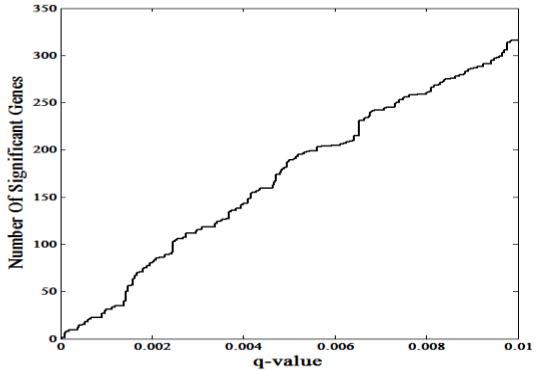
Histogram of t-test results





p-value versus t-statistics and q-value versus t-statistics plots.





Plot of number of significant genes filtered for respective p-value and q-value cutoff

Result

- Number of genes filtered for 4 statistically significant p and q value cutoffs

Cut off value	No. of Significant genes based on p-value	No. of Significant genes based on q-value
0.001	374	31
0.005	968	168
0.01	1231	316
0.05	2659	1270

- We use q-value cutoff of 0.005 and the resultant list of 168 genes is identified as differentially expressed between the diseased and healthy population.

Clustering of Significant Genes

Objective

- Clustering - statistical technique used to generate a category structure that fits a set of observations.
- High degree of association between members of the same group, low degree of association between members of different groups.
- Clustering of gene expression patterns is used to identify groups of co-expressed genes.
- Genes belonging to same cluster are typically involved in related functions and are frequently co-regulated..

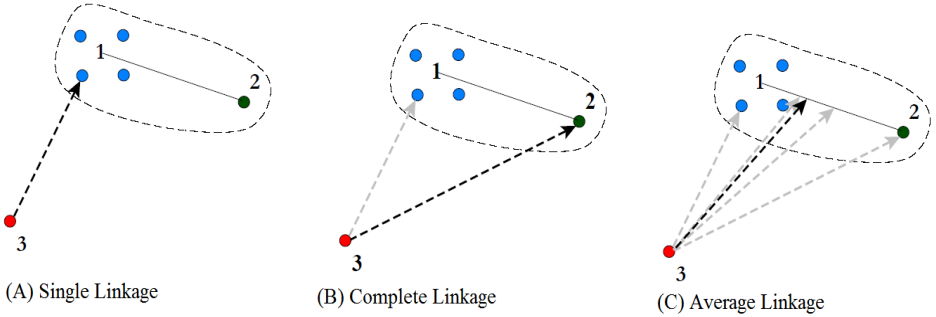
Methodology

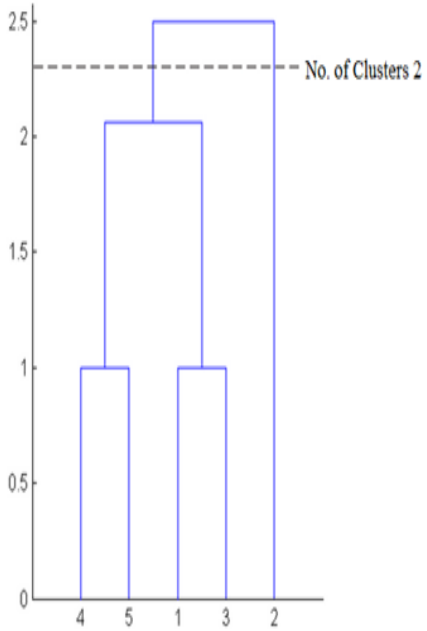
- Hierarchical clustering: agglomerative

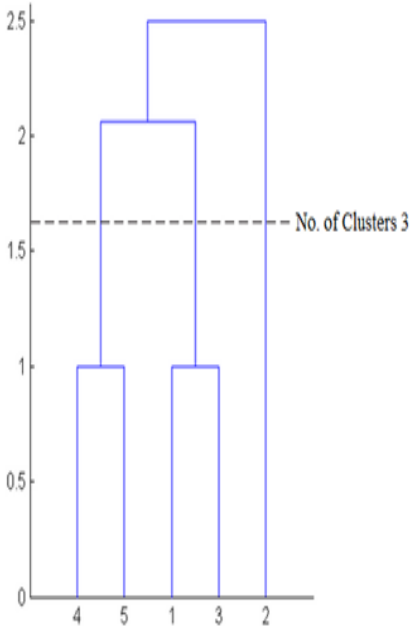
Distance Measures

- Euclidean Distance $D_{Euc}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Manhattan Distance $D_{Man}(X,Y) = \sum_{i=1}^{n} | x_i - y_i |$
- Mahalanobis distance $Dis_{Mah} = \sqrt{(X-Y)C^{-1}(X-Y)^T}$ where $C_{ij} = \mathrm{cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$
- Pearson correlation coefficient $Dis_{Corr} = 1 - \frac{(X-\overline{X})(Y-\overline{Y})^T}{\sqrt{(X-\overline{X})(X-\overline{X})^T}\sqrt{(Y-\overline{Y})(Y-\overline{Y})^T}}$
- Spearman Rank correlation coefficient $Dis_{Spr} = 1 - \frac{(R_X-\overline{R_X})(R_Y-\overline{R_Y})^T}{\sqrt{(R_X-\overline{R_X})(R_X-\overline{R_X})^T}\sqrt{(R_Y-\overline{R_Y})(R_Y-\overline{R_Y})^T}}$

Linkage criterion and Dendrogram







- **Dendrogram :** hierarchical series of nested clusters graphically represented by a tree, called dendrogram.
- Cutting the dendrogram at different depths produces different number of clusters

Cluster Validity

Hierarchical Clustering

Cophenetic Correlation Coefficient (CPCC) measures original pairwise dissimilarities between the feature vectors and the cophenetic dissimilarities from the dendrogram

Partitional Clustering

- Dunn's Index
- Davies-Bouldin Index
- Calinski- Harabasz Index
- Silhouette Index
- CS Index
- SYM-Index
- SV-Index

Distance and Linkage criterion Selection

Hierarchical Cluster Analysis on Healthy Population

	Single Linkage	Average Linkage	Complete Linkage
Euclidean Distance	0.7024	0. 8263	0.7483
Manhattan Distance	0.7022	0.8457	0.7586
Mahalanobis Distance	0.8205	0.8437	0.3039
Pearson Correlation Distance	0.7071	0.8315	0.8128
Spearman Correlation Distance	0.6750	**0.8484**	0.7612

Hierarchical Cluster Analysis on Diseased Population

	Single Linkage	Average Linkage	Complete Linkage
Euclidean Distance	0.6515	0.8254	0.7230
Manhattan Distance	0.6669	0.8172	0.7182
Mahalanobis Distance	0.8319	0.8550	0.3216
Pearson Correlation Distance	0.8844	0.9118	0.8653
Spearman Correlation Distance	0.9046	**0.9224**	0.8843

Deciding the k - parameter

For Healthy Population

k-parameter	Dunn's Index	Davies-Bouldin Index	Calinski-Harabasz Index	Silhouette Index	CS Index	SYM Index	SV Index
2	0.519489466	**0.421794**	179.4794	0.320344	2.417697	0.123279	0.058357
3	**0.521822826**	1.285558	**240.179**	**0.341226**	3.007731	**0.202354**	**0.475018**
4	0.301566338	1.354472	127.1219	0.267722	3.691343	0.083151	0.375878
5	0.301566338	1.235196	96.22031	0.249841	2.543032	0.11748	0.397804
6	0.301566338	1.175929	77.90129	0.233563	**2.130993**	0.187262	0.417877
7	0.301566338	1.100681	65.56314	0.226116	2.929643	0.169686	0.275173
8	0.301566338	1.026924	56.36059	0.214371	4.345251	0.150266	0.187871
9	0.250779006	1.204307	51.56966	0.22683	5.736134	0.133218	0.151053
10	0.250779006	1.215522	46.92917	0.225899	6.390402	0.122311	0.132319

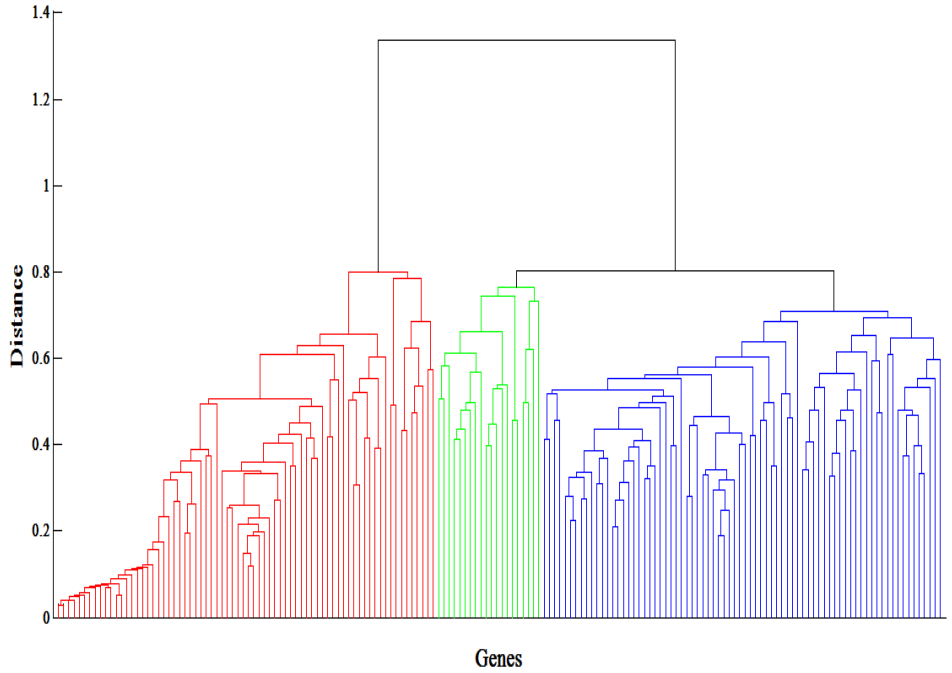
Deciding the k - parameter

For Diseased Population

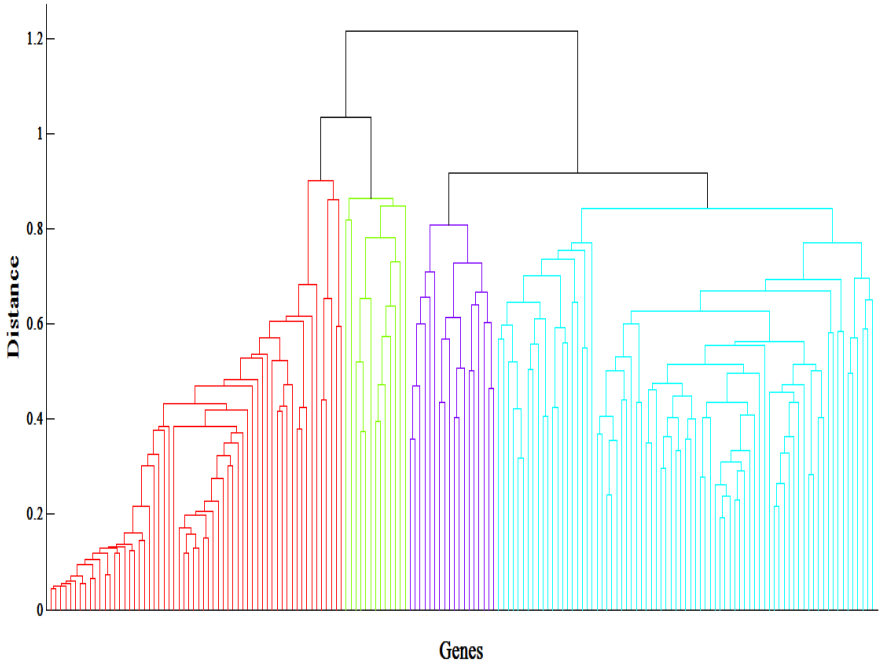
k-parameter	Dunn's Index	Davies-Bouldin Index	Calinski-Harabasz Index	Silhouette Index	CS Index	SYM Index	SV Index
2	0.323957822	**0.546018**	102.3888	0.3014	2.51132	0.0931	0.093128
3	0.339238222	0.900783	137.9075	0.3158	2.472717	0.1532	0.153163
4	**0.375366245**	1.121942	**240.9651**	**0.3796**	1.976578	**0.2221**	0.122059
5	0.339434355	1.061795	80.1508	0.2891	5.670765	0.0992	0.099245
6	0.296278403	1.031703	64.8691	0.2763	6.60191	0.1524	0.152398
7	0.296278403	0.965214	54.5249	0.2595	1.219989	0.1577	0.157732
8	0.296278403	0.883861	47.1872	0.2603	**1.067121**	0.1773	**0.177312**
9	0.352699333	0.936416	44.0637	0.2698	9.959585	0.1571	0.157077
10	0.352699333	0.850708	39.3389	0.274	8.445925	0.147	0.147001

Dendrogram for optimal k-parameter

For Healthy Population



For Diseased Population



Cluster Assignment Analysis

- For healthy population genes distributed as

Cluster#**1** 20 genes

Cluster#**2** 72 genes

Cluster#**3** 76 genes

- For diseased population genes distributed as

Cluster#**1** 13 genes

Cluster#**2** 18 genes

Cluster#**3** 60 genes

Cluster#**4** 77 genes

- Cluster# 1 of Healthy Dataset and Cluster# 2 of Diseased Dataset 72.22% similarity
- Cluster# 2 of Healthy Dataset and Cluster# 3 of Diseased Dataset 100% similarity
- Cluster# 3 of Healthy Dataset and Cluster# 4 of Diseased Dataset 89.61% similarity
- Extra cluster of diseased population: 10 genes (Cluster# 2) , 2 genes (Cluster# 1) and 1 gene (Cluster# 3) of healthy dataset
- We identify 12 genes showing different cluster assignments in healthy and diseased population.

Biological Enrichment of Clusters

- Gene Ontology
- Provides descriptions about gene products in terms of

  biological process (BP)

  molecular function (MF)

  cellular components (CC)

- GO Term Finder tool (http://go.princeton.edu/)
- Significant GO terms shared between queried genes
- p-value : degree of enrichment
- Small p-value – stronger evidence of annotation

Healthy Population Cluster Analysis

Cluster#1 : **regulatory cluster**

GOID	GO TERM from biological_process Ontology	p-value < 0.05	% of Genes of Cluster Annotated
GO:0044237	cellular metabolic process	0.004534	75
GO:0051128	regulation of cellular component organization	0.000635	30
GO:0048583	regulation of response to stimulus	0.014725	30
GO:0022414	reproductive process	0.000166	25
GO:0009653	anatomical structure morphogenesis	0.008663	25
GO:0080134	regulation of response to stress	0.002823	25
GO:0009719	response to endogenous stimulus	0.005762	20
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.019413	20
GO:0002697	regulation of immune effector process	0.000658	15
GO:0050679	positive regulation of epithelial cell proliferation	0.001952	10

Healthy Population Cluster Analysis

Cluster#2 : **response and signaling cluster**

GOID	GO TERM from biological_process Ontology	p-value <0.001	% of Genes of Cluster Annotated
GO:0050896	response to stimulus	1.42E-06	65.55556
GO:0007154	cell communication	8.44E-05	53.05556
GO:0044700	single organism signaling	0.00014	41.66667
GO:0007165	signal transduction	0.000218	38.88889
GO:0044707	single-multicellular organism process	1.64E-05	37.5
GO:0048518	positive regulation of biological process	5.19E-06	36.11111
GO:0010468	regulation of gene expression	0.000363	31.94444
GO:0044767	single-organism developmental process	0.00053	30.55556
GO:0006950	response to stress	4.40E-05	29.16667
GO:0002376	immune system process	2.78E-05	22.22222
GO:0010646	regulation of cell communication	0.0004	22.22222
GO:0035556	intracellular signal transduction	0.000794	22.22222
GO:0006955	immune response	2.41E-05	18.05556
GO:0006952	defense response	0.000126	16.66667

Healthy Population Cluster Analysis

Cluster#3 : **cell development and maintenance cluster**

GOID	GO TERM from biological_process Ontology	p-value <0.001	% of Genes of Cluster Annotated
GO:0019222	regulation of metabolic process	0.000159	43.42105
GO:0044767	single-organism developmental process	6.34E-06	35.52632
GO:0048856	anatomical structure development	6.93E-06	32.89474
GO:0030154	cell differentiation	0.000472	32.36842
GO:0048513	organ development	6.15E-05	21.05263
GO:0008283	cell proliferation	0.000142	15.78947
GO:0008219	cell death	0.000648	15.78947
GO:0010941	regulation of cell death	0.00099	13.15789
GO:0043687	post-translational protein modification	2.64E-06	9.210526
GO:0006281	DNA repair	0.000851	9.210526
GO:0044770	cell cycle phase transition	3.33E-05	9.210526
GO:0010564	regulation of cell cycle process	6.58E-05	9.210526
GO:1901214	regulation of neuron death	0.000786	5.263158

CONSTRUCTION OF GENE REGULATORY NETWORK

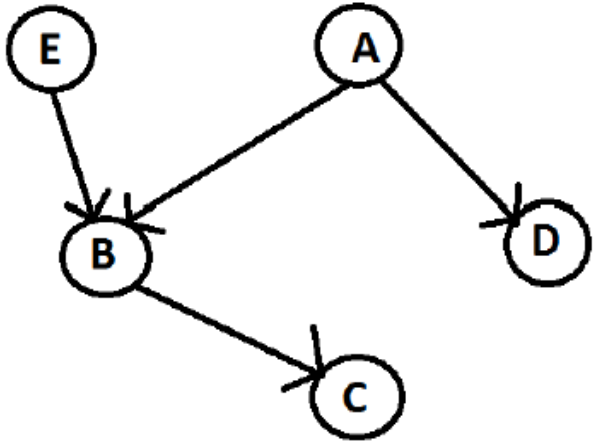
Objective

- GRN - most important organizational level in the cell, signals from the cell state and the outside environment are integrated in terms of activation and inhibition of genes.
- "**Connection**" connotes regulatory interaction
- Genes having similar gene expression profiles are more likely to regulate one another or be regulated by some other common parent gene.
- Bayesian Network approach
- Modified version of Sparse Candidate Algorithm [6]

Bayesian Network Formalisms

- Represents joint probability distribution of random variables consisting of two components $B =< \mathbf{G}, \varphi >$
- $\mathbf{G}$ -DAG vertices correspond to $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$
- Encodes Markov assumption $\forall X_i \ \left(X_i \perp Non-Descendants_{X_i} \mid \mathbf{U}_{X_i}\right)$
- φ conditional probability distribution (CPD) of $P\left(X_i \mid \mathbf{U}_{X_i}\right)$
- Joint distribution satisfying conditional independence properties decomposed by chain rule for Bayesian networks

$$P\left(X_1, X_2, \cdots, X_N\right) = \prod_{i=1}^{N} P\left(X_i \mid \mathbf{U}_{X_i}\right)$$



$(A \perp E), (B \perp D \mid A, D), (C \perp A, D, E \mid B)$

$(D \perp B, C, E \mid A) \text{ and } (E \perp A, D)$

$$P(A, B, C, D, E) = P(A)P(E)P(\mathrm{B} \mid A, E)P(C \mid B)P(D \mid A)$$

Learning Bayesian Network

- Given a training set of samples $\boldsymbol{D} = \mathrm{x}[1], \mathrm{x}[2], \cdots, \mathrm{x}[M]$ independently drawn from some unknown Bayesian Network $\mathbf{G}^*$ with underlying distribution $\mathbf{P}^*$, goal : recover $\mathbf{G}^*$
- **Parameter Learning** : given network structure $\mathbf{G}$, set of data instances $\boldsymbol{D}$ determine what values of $\boldsymbol{\varphi}$ best describes the generated data.
- Maximum Likelihood estimation : $L(\boldsymbol{\varphi} : \boldsymbol{D}) = \prod_{m=1}^{M} P(\mathrm{x}[m] \mid \boldsymbol{\varphi})$,choose $\bar{\boldsymbol{\varphi}} = \max_{\boldsymbol{\varphi}} L(\boldsymbol{\varphi} : \boldsymbol{D})$ closed form : $\bar{\boldsymbol{\varphi}} = \frac{\mathrm{M}[x, \mathrm{u}]}{\mathrm{M}[\mathrm{u}]}$
- **Structure Learning** :define scoring function that measures how well each model fits the observed data. Optimization algorithm employed to search for highest scoring model.
- By Bayes rule, $P(\mathbf{G} \mid \boldsymbol{D}) = \frac{P(\boldsymbol{D} \mid \mathbf{G}) P(\mathbf{G})}{P(\boldsymbol{D})}$ ignoring normalization factor $score_B(\mathbf{G} : \boldsymbol{D}) = \log P(\boldsymbol{D} \mid \mathbf{G}) + \log P(\mathbf{G})$

where $P(\boldsymbol{D} \mid \mathbf{G}) = \int P(\boldsymbol{D} \mid \mathbf{G}, \boldsymbol{\varphi}) P(\boldsymbol{\varphi} \mid \mathbf{G}) \, d\boldsymbol{\varphi}$ and by decomposability $score_B(\mathbf{G} : \boldsymbol{D}) = \sum_{i=1}^{N} Score\_Contribution\left(X_i, \mathrm{U}_{X_i}^{\mathbf{G}} : \boldsymbol{D}\right)$

Search Algorithm

- **Input** : training set , scoring function, set of possible network structures
- **Output** : network structure that maximizes score
- $2^{O(n^2)}$ possible structures : NP-Hard problem
- Initial Network Formation
- Discretization : **Z-score**
- Mutual Information $\mathrm{MI}(X_i, X_j) = \sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$
- Select k nodes giving highest MI as candidate parents for X_i

Cycle_Removal (B_n)

Repeat until B_n is not a DAG

Detect cycle in graph using DFS : $C = X_{i1} \to X_{i2} \to \cdots \to X_{im} \to X_{i1}$

List nodes involved in cycle $L = \{X_{i1}, X_{i2}, \cdots, X_{im}\}$

$\forall X_{ij} \in L$ find $parent(X_{ij})$ in C

Find rank of $parent(X_{ij})$ as a candidate parent of X_{ij}

Construct $R = \{r_{i1}, r_{i2}, \cdots, r_{im} : r_{ij} = rank(parent(X_{ij})) \text{ in } C\}$

Find $r_{iq} = \max\{R\}$ and $X_t = parent(X_{iq})$ in C

Remove edge $X_t \to X_{iq}$ and update $B_n = B_n - edge(X_t \to X_{iq})$

Return B_n

Sparse Candidate Algorithm

Input:

- The training Dataset $\boldsymbol{D} = \{x[1], x[2], \cdots, x[M]\}$
- An initial network $\boldsymbol{B_0}$
- A decomposable score such that $score(\boldsymbol{B} \mid \boldsymbol{D}) = \sum_{i=1}^{N} Score\_Contribution(X_i, U_{X_i}^{B} : \boldsymbol{D})$
- A parameter $\boldsymbol{k}$ = maximum parent for each node
- A parameter $\boldsymbol{l}$ = minimum number of edges in network

Output: A network B

Loop for $n = 1, 2, \cdots$ until convergence

Restrict:

Based on $\boldsymbol{D}$ and $\boldsymbol{B_{n-1}}$ for each variable select a set C_i^n of candidate parents with $|C_i^n| \leq \boldsymbol{k}$ by ***Candidate_Parent_Selection*** $(X_i, \boldsymbol{B_{n-1}}, \boldsymbol{D}, \boldsymbol{k})$

This defines a directed graph $\boldsymbol{H_n} = (\mathbf{X}, E)$ where $E = \{X_j \rightarrow X_i \mid \forall i, j \; X_j \in C_i^n\}$

Remove cycles in $\boldsymbol{H_n}$ using ***Cycle_Removal*** $(\boldsymbol{H_n})$

Maximize:

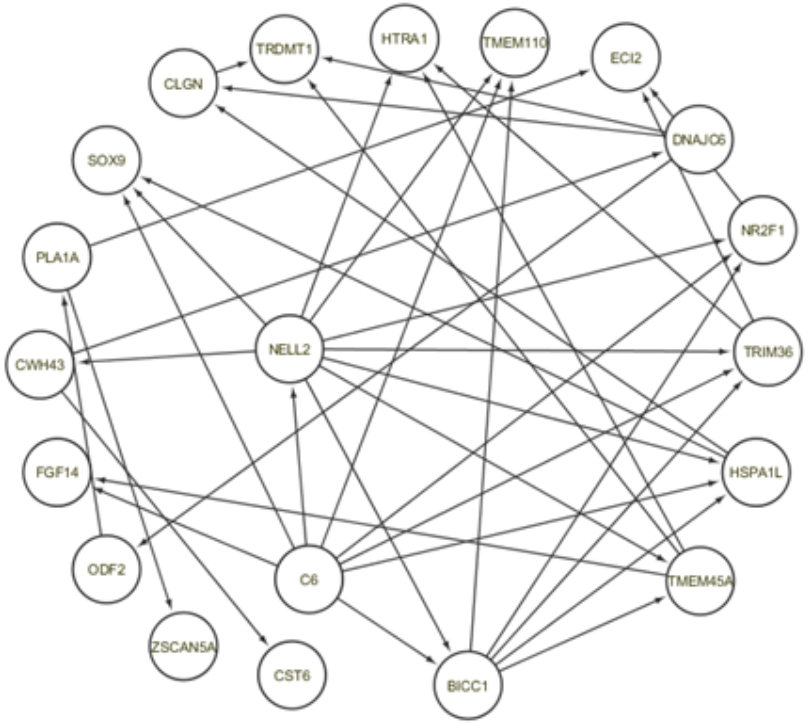
Find network $\boldsymbol{B_n} = \langle \mathbf{G_n}, \boldsymbol{\varphi} \rangle$ maximizing $score(\boldsymbol{B_n} \mid \boldsymbol{D})$ among networks satisfying $\mathbf{G_n} \subset \boldsymbol{H_n}$ (i.e. $\forall X_i, \; U_{X_i}^{G_n} \subseteq C_i^n$) using ***Greedy_Hill_Climbing***.

Restrict the minimum number of edges by $\boldsymbol{l}$ to avoid sparse graph

Return $\boldsymbol{B_n}$

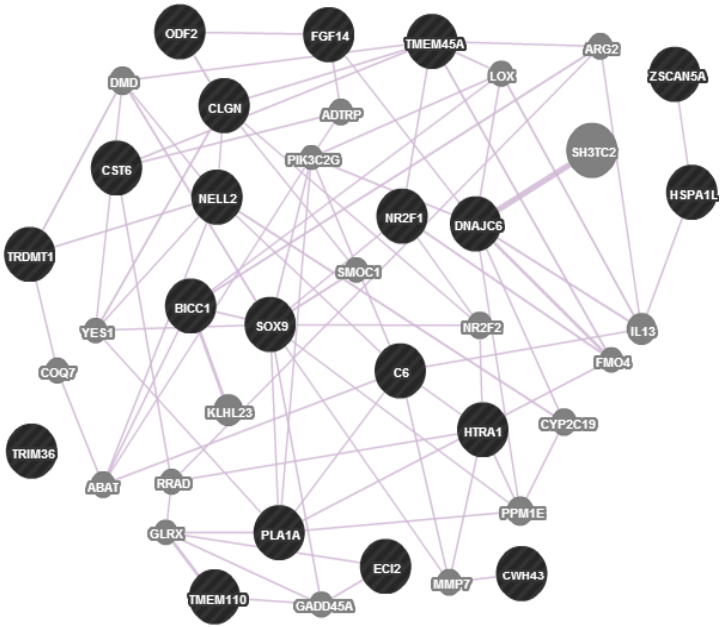
Results

Initial network based on mutual information between genes, 60 edges



Final GRN for Cluster# 1 of healthy population, 39 edges

Validation of predicted network



Gene Interaction Network returned by GeneMANIA

Serial #	GeneMANIA Network Edge or Path	Network Study for the edge	Bayesian Network Edge Connection	Connectivity between nodes in Bayesian network	Dependence Evident From
1	C6 – NELL2	Ramaswamy-Golub-2001	$C6 \rightarrow NELL2$	Parent-Child	Causal Reasoning
2	C6 – HTRA1	Burington-Shaughnessy-2008	$C6 \rightarrow NELL2 \rightarrow HTRA1$	Descendant	Causal flow of dependence
3	BICC1–SOX9	Burington-Shaughnessy-2008; Innocenti-Brown-2011	C6 ↙ SOX9, C6 ↘ BICC1	Common Parent	Evidential reasoning causal reasoning
4	BICC1 –SOX9 – NR2F1	Kang-Willman-2010	$BICC1 \rightarrow NR2F1$	Parent-Child	Causal Reasoning
5	CLGN –ODF2	Mallon-McKay-2013	DNAJC6 ↙ ODF2, DNAJC6 ↘ CLGN	Common Parent	Evidential reasoning causal reasoning

Conclusion

- Genome wide expression profile analysis of healthy and diseased population
- Identification of genes exhibiting differential expression pattern between healthy and diseased population
- Clustering genes into groups of co-expressed genes
- Construction of intra-cluster gene regulatory network

Future Scope

- Incorporation of fuzzy clustering measures
- Model GRN using continuous variables Bayesian network
- Inter-Cluster GRN with efficient learning algorithms
- Analysis of network connectivity change between healthy and diseased population

References

[1] Spira A, Beane JE, Shah V, Steiling K et al. “Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer”. Nat Med 2007 Mar;13(3). [PMID: 17334370]

[2] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771

[3] Storey, J.D., and Tibshirani, R.(2003), “Statistical significance for genomewide studies”, PNAS 100(16), pp 9440–9445.

[4] Daphne Koller and Nir Friedman, *Probabilistic Graphical Models : Principles and Techniques*. MIT Press, 2009.

[5] Nir Friedman, Michal Linial, Iftach Nachman and Dana Pe'er, “Using Bayesian Networks to Analyze Expression Data”, Journal of Computational Biology. August 2000, 7(3-4), pp 601-620.

[6] N. Friedman, I. Nachman, and D. Pe’er. “Learning Bayesian network structure from massive datasets: The “sparse candidate” algorithm”. In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI), pp 196–205. 1999.

THANK YOU