

2009

## Identifying prognostic gene-signatures using a network-based approach

Swetha Bose Nutakki  
*West Virginia University*

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>

---

### Recommended Citation

Nutakki, Swetha Bose, "Identifying prognostic gene-signatures using a network-based approach" (2009). *Graduate Theses, Dissertations, and Problem Reports*. 4509.  
<https://researchrepository.wvu.edu/etd/4509>

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact [researchrepository@mail.wvu.edu](mailto:researchrepository@mail.wvu.edu).

**IDENTIFYING PROGNOSTIC GENE-SIGNATURES USING A  
NETWORK-BASED APPROACH**

Swetha Bose Nutakki

Thesis Submitted to the  
College of Engineering and Mineral Resources  
At West Virginia University  
In partial fulfillment of the requirements  
For the degree of

Master of Science  
In  
Electrical Engineering

Committee:

Nancy Lan Guo, Ph.D., Advisor/Chair  
Bojan Cukic, Ph.D., (Departmental Chair)  
Arun A. Ross, Ph.D.,  
James Denvir, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, WV

2009

Keywords: Implication Network, Prediction Logic, Bayesian Network, Protein-Protein Interactions, Poor & Good Prognosis, Differential Components, Cancer Hallmarks, Prognostic & Topological Validation, Time dependent ROC analysis, Random Test, Cox Proportional Hazard Model, CPE, GSEA, PRODISTIN, TETRAD IV, NCI Pathways, PubMed, MATISSE, KEGG, STRING 8, Ingenuity Pathway Analysis, Pathway Studio

## Abstract

### Identifying Prognostic Gene Signatures Using a Network-Based Approach Swetha Bose Nutakki

The main objective of this study is to develop a novel network-based methodology to identify prognostic signatures of genes that can predict recurrence in cancer. Feature selection algorithms were used widely for the identification of gene signatures in genome-wide association studies. But most of them do not discover the causal relationships between the features and need to compromise between accuracy and complexity. The network-based techniques take the molecular interactions between pairs of genes into account and are thus a more efficient means of finding gene signatures, and they are also better in terms of its classification accuracy without compromising over complexity. Nevertheless, the network-based techniques currently being used have a few limitations each. Correlation-based coexpression networks do not provide predictive structure or causal relations among the genes. Bayesian networks cannot model feedback loops. Boolean networks can model small scale molecular networks, but not at the genome-scale. Thus the prediction logic induced implication networks are chosen to generate genome-wide coexpression networks, as they integrate formal logic and statistics and also overcome the limitations of other network-based techniques.

The first part of the study includes building of an implication network and identification of a set of genes that could form a prognostic signature. The data used consisted of 442 samples taken from 4 different sources. The data was split into training set UM/HLM ( $n=256$ ) and two testing sets DFCI ( $n=82$ ) and MSK ( $n=104$ ). The training set was used for the generation of the implication network and eventually the identification of the prognostic signature. The test sets were used for validating the obtained signature. The implication networks were built by using the gene expression data associated with two disease states (metastasis or non-metastasis), defined by the period and status of post-operative survival. The gene interactions that differentiated the two disease states, the differential components, were identified. The major cancer hallmarks (E2F, EGF, EGFR, KRAS, MET, RB1, and TP53) were considered, and the genes that interacted with all the major hallmarks were identified from the differential components to form a 31-gene prognostic signature. A software package was created in R to automate this process which has C-code embedded into it. Next, the signature was fitted into a COX proportional hazard model and the nearest point to the perfect classification in the ROC curve was identified as the best scheme for patient stratification on the training set (log-rank  $p$ -value =  $1.97e-08$ ), and two test sets DFCI (log-rank  $p$ -value =  $2.13e-05$ ) and MSK (log-rank  $p$ -value =  $1.24e-04$ ) in Kaplan-Meier analyses.

Prognostic validation was carried out on the test sets using methods such as Concordance Probability Estimate (CPE) and Gene Set Enrichment Analysis (GSEA). The accuracy of this signature was evaluated with CPE, which achieves 0.71 on the test set DFCI (log-rank  $p$ -value =  $5.3e-08$ ) and 0.70 on test set MSK (log-rank  $p$ -value =  $2.1e-07$ ). The hazard ratio of this 31-gene prognostic signature is 2.68 (95% CI: [1.88, 3.82]) on the DFCI dataset and 3.31 (95% CI: [2.11, 5.2]) on the MSK set. These results demonstrate that our 31-gene signature was significantly more accurate than previously published signatures on the same datasets. The false discovery rate (FDR) of this 31-gene signature is 0.21 as computed with GSEA, which showed that our 31 gene signature was comparable to other lung cancer prognostic signatures on the same datasets.

Topological validation was performed on the test sets for the identified signature to validate the computationally derived molecular interactions. The interactions from implication networks were compared with those from Bayesian networks implemented in Tetrad IV. Various curated databases and bioinformatics tools were used in the topological evaluation, including PRODISTIN, KEGG, PubMed, NCI-Nature pathways, MATISSE, STRING 8, Ingenuity Pathway Analysis, and Pathway Studio 6. The results showed that the implication networks generated all the curated interactions from various tools and databases, whereas Bayesian networks contained only a few of them. It can thus be concluded that implication networks are capable of generating many more gene or protein interactions when compared to the currently used network techniques such as Bayesian networks.

## **Acknowledgements**

I would like to thank Dr. Nancy Lan Guo for giving me an opportunity to work on the current project. I would like to thank her for the financial assistance I received throughout the three years I spent in this lab. I would like to thank her for being my advisor and supporting me all through my study. This project has been supported by NIH/R01LM009500, NIH/NCRR P20RR16440, and NIH/NCRR P2016477.

I would like to thank Dr. Bojan Cukic for being the Chairperson of my committee. I would like to thank Dr. Arun Ross for being a part of my committee. I would like to thank Dr. Jim Denvir for his help in making me understand the statistical analyses. I would like to thank Dr. Yong Qian from NIOSH for the help in biological interpretation of the 31 genes in the signature. I would like to thank Dr. Valenti for his support and encouragement.

I would like to thank all my lab mates Ying-Wooi Wan, Jason Young, Kursad Tosun, Shruthi Rathnagiriswaran, Ramakanth Reddy Mettu, Joseph Putila, Ebrahim Sabbagh, and Jia Jia Wang for their support throughout and for patiently listening to my results every week. I would like to thank Jason Young who worked a major part in introducing the project to me when I entered the lab and then helped me in creating a code compatible to the package. I would like to thank Ying-Wooi Wan in particular for her help in all the queries I had and for her timely encouragement and support.

I would like to thank my friends and family members who gave me a lot of moral support throughout my Masters.

## Contents

Abstract.....	ii
Acknowledgements.....	iii
List of Figures.....	vii
List of Tables.....	xiii
1 Introduction.....	1
2 Background.....	7
2.1 Introduction.....	7
2.2 Different techniques to identify signatures.....	7
2.2.1 Coexpression Network.....	9
2.2.2 Bayesian Network.....	13
2.2.3 Artificial Neural Network.....	21
2.2.4 Boolean Networks.....	25
2.3 Implication Networks.....	26
2.4 Survival Analysis.....	32
2.4.1 Time dependent ROC analysis and Random Test.....	33
2.4.2 COX Proportional Hazards model.....	34
2.4.3 Multivariate Analysis using COX Proportional Hazards model.....	35
2.4.4 Kaplan-Meier Plots.....	37
2.5 Prognostic Validation.....	37
2.5.1 Overall Accuracy.....	38
2.5.2 Concordance Probability Estimate (CPE).....	39
2.5.3 Gene Set Enrichment Analysis (GSEA).....	41
2.6 Topological Validation.....	42
2.6.1 PRODISTIN.....	42
2.6.2 PubMed.....	45
2.6.3 NCI Pathways (Pathway Interaction Database).....	47
2.6.4 KEGG.....	49
2.6.5 MATISSE.....	52
2.6.6 STRING 8.....	54

2.6.7	Ingenuity Pathway Analysis .....	57
2.6.8	Pathway Studio .....	59
2.7	Summary .....	61
3	Network-based approach for identification of prognostic signatures.....	63
3.1.	Introduction .....	63
3.2.	Methodology for identifying prognostic gene signatures .....	64
3.2.1	Datasets Information .....	65
3.2.2	Dataset processing .....	65
3.2.3	Deriving genome scale gene interactions .....	66
3.2.4	Identifying Differential Components .....	67
3.2.5	Major Cancer Hallmarks used to identify prognostic markers .....	68
3.2.6	Identifying Gene Signatures .....	68
3.3.	Survival Analysis .....	72
3.3.1	Time dependant ROC analysis and Random Test .....	72
3.3.2	Cox proportional hazards model on 11, 21 and 31 Gene Signatures .....	75
3.4.	Summary .....	83
4	Prognostic Validation, Clinical Evaluation & Topological Validation.....	84
4.1	Introduction .....	84
4.2	Gene Signature Details and Differentially Expressed genes.....	84
4.3	Prognostic Validation.....	92
4.3.1	Overall Accuracy:.....	92
4.3.2	Concordance Probability Estimate (CPE) .....	95
4.3.3	Gene Set Enrichment Analysis(GSEA).....	97
4.3.4	Comparison of model with other classification methods using WEKA.....	101
4.4	Clinical Evaluation.....	108
4.5	Topological Validation.....	111
4.5.1	PRODISTIN.....	117
4.5.2	TETRAD IV.....	132

4.5.3	KEGG .....	137
4.5.4	NCI Pathways .....	138
4.5.5	PubMed interactions .....	139
4.5.6	Matisse .....	140
4.5.7	STRING 8 .....	141
4.5.8	Ingenuity Pathway Analysis .....	143
4.5.9	Pathway Studio .....	147
4.6	Summary .....	148
5	Software Implementation .....	154
5.1	Introduction .....	154
5.2	Description .....	154
5.3	Results & Screenshots .....	156
5.4	Summary .....	158
6	Conclusions & Prospective Work.....	159
6.1	Conclusions .....	159
6.2	Future Work .....	161
	List of References .....	162

## List of Figures

Figure 2-1: Model used in TETRAD IV showing all the boxes .....	16
Figure 2-2: Data Wrapper shown.....	16
Figure 2-3: Available Search Algorithms .....	17
Figure 2-4: PC Search after execution .....	17
Figure 2-5: Bayes Parametric Model .....	18
Figure 2-6: ML Bayes Estimator .....	19
Figure 2-7: Bayes Instantiated Model.....	20
Figure 2-8: ROC curve after classification .....	21
Figure 2-9: Confusion Matrix after Classification.....	21
Figure 2-10: Structural diagram of a general artificial neural network .....	22
Figure 2-11: Training and Prediction Phases in a model built with Artificial Neural Networks [10].....	23
Figure 2-12: Six possible implications relating two variables [2] .....	28
Figure 2-13: Contingency Table .....	29
Figure 2-14: Implication induction algorithm from Guo et al [2] .....	30
Figure 2-15: Step wise procedure for PRODISTIN method [21].....	44
Figure 2-16: PRODISTIN website.....	45
Figure 2-17: PubMed website.....	46
Figure 2-18: Advanced Search in PubMed.....	46
Figure 2-19: Pathway Interaction Database.....	48
Figure 2-20: Browsing Pathways in PID .....	48
Figure 2-21: KEGG Website .....	51
Figure 2-22: KEGG PATHWAY.....	51
Figure 2-23: MATISSE interface.....	52
Figure 2-24: Choosing the Algorithm.....	52
Figure 2-25: Displaying the module .....	53



Figure 2-26: Module from MATISSE .....	54
Figure 2-27: Module from Co-clustering method.....	54
Figure 2-28: STRING 8 web interface.....	56
Figure 2-29: Different Network and Path Designer shapes along with Relationships used in IPA to represent different types of data.....	59
Figure 2-30: Different shapes and colors used in Pathway Studio to represent for different types of data.....	61
Figure 3-1: Flow chart of the methodology .....	63
Figure 3-2: Bar graph showing the number of interactions in Poor (high risk) and Good (low risk) prognosis for genome wide interactions and differential components in the Training dataset .....	67
Figure 3-3: Finding Nearest point as cutoff for Training data for 31 genes .....	76
Figure 3-4: KM plot of Training data with nearest point cutoff for 31 genes .....	76
Figure 3-5: KM plot of DFCI data with nearest point cutoff of training data for 31 gene signature .....	77
Figure 3-6: KM plot of MSK data with nearest point cutoff of training data for 31 gene signature .....	77
Figure 3-7: KM plot of Training data for Stage I patients .....	79
Figure 3-8: KM plot of DFCI data for Stage I patients.....	80
Figure 3-9: KM plot of MSK data for Stage I patients .....	80
Figure 3-10: KM plot of Training data for Stage IA patients .....	81
Figure 3-11: KM plot of DFCI data for Stage IA patients.....	82
Figure 3-12: KM plot of MSK data for Stage IA patients .....	82
Figure 3-13: KM plot of Training data for Stage IB patients .....	82
Figure 3-14: KM plot of DFCI data for Stage IB patients .....	83
Figure 3-15: KM plot of MSK data for Stage IB patients .....	83
Figure 4-1: Histograms for 4 genes over all the 442 samples of data showing that the log transformed data is less skewed than data which was not log transformed.....	86
Figure 4-2: Histograms for 4 samples, each from one dataset over all the 22215 genes of data showing that the log transformed data is less skewed than data which was not log transformed .....	86

Figure 4-3: Fold changes of the 31 genes for Stage, where blue color bars represent fold change of stage 2 w.r.t. stage 1 and red color bars represent fold change of stage 3 w.r.t. stage 1 and genes with stars on the top represent the significant genes from T-test with $p \leq 0.05$ .....	91
Figure 4-4: Fold changes of the 31 genes for Tumor Differentiation, where blue color bars represent fold change of Moderate differentiation w.r.t. Well differentiation and red color bars represent fold change of Poor differentiation w.r.t. Well differentiation and genes with stars on the top represent the significant genes from T-test with $p \leq 0.05$ .....	91
Figure 4-5: Fold changes of the 31 genes for Lymph node metastases, where blue color bars represent fold change of lymph node positive w.r.t. lymph node negative. ....	92
Figure 4-6: Hazard ratios and 95% Confidence Intervals (obtained from the CPE package which use the risk scores of the entire signature as input) shown along with error bars for 31gene signature and the model from Shedden et al. ....	96
Figure 4-7: Comparison of p-values (obtained from the CPE package which use the risk scores of the entire signature as input) for 31 gene signature and model from Shedden et al. on a logarithmic scale.....	96
Figure 4-8 : Concordance Probability Estimates compared between 31 gene signature and the model from Shedden et al. ....	97
Figure 4-9: Screenshot for loading data in to GSEA .....	98
Figure 4-10: Screenshot showing the Basic fields in running GSEA.....	99
Figure 4-11: Screenshot showing the Selection of Metric for ranking genes.....	99
Figure 4-12: Screenshot showing the selection of sorting the gene list based on their real values .....	100
Figure 4-13: Enrichment score plot for the 31 gene signature picked from implication networks which shows the Enrichment profile on the top and the ranked list metric on the bottom.....	101
Figure 4-14: Plot showing the Nominal Enrichment Scores, False Discovery Rates and Nominal P-values for all the signatures with Signature index of each signature from Table 4-9 and 4-10. Index 4 represents the 31-gene signature from implication networks .....	101
Figure 4-15: Differential Components common to Train & DFCI datasets in high risk group..	115
Figure 4-16: Differential Components common to Train & MSK datasets in high risk group ..	115
Figure 4-17: Differential Components common to DFCI & MSK datasets in high risk group..	115
Figure 4-18: Differential Components common to all 3 datasets in high risk group .....	115
Figure 4-19: Differential Components common to Train & DFCI datasets in low risk group...	116
Figure 4-20: Differential Components common to Train & MSK datasets in low risk group ...	116

Figure 4-21: Differential Components common to DFCI & MSK datasets in low risk group...	116
Figure 4-22: Differential Components common to all three datasets in low risk group.....	116
Figure 4-23: Differential Components common among the three datasets in both the prognosis groups where good prognosis corresponds to low risk group and poor prognosis corresponds to high risk group and the major molecular and cellular functions identified from IPA were also shown .....	117
Figure 4-24: Clustering from interactions Common to Train and DFCI Metastasis group from PRODISTIN.....	119
Figure 4-25: Clustering from interactions Common to Train and MSK Metastasis group from PRODISTIN.....	121
Figure 4-26: Clustering from interactions Common to DFCI and MSK Metastasis group from PRODISTIN.....	123
Figure 4-27: Clustering from interactions Common to Train &DFCI Non-Metastasis group from PRODISTIN.....	124
Figure 4-28: Clustering from interactions Common to Train & MSK Non-Metastasis group from PRODISTIN.....	126
Figure 4-29: Clustering from interactions Common to DFCI & MSK Non-Metastasis group from PRODISTIN.....	128
Figure 4-30: Clustering from interactions Common to 3 datasets Metastasis group from PRODISTIN.....	130
Figure 4-31: Clustering from interactions Common to 3 datasets Non-Metastasis group from PRODISTIN.....	131
Figure 4-32: Model used to build Bayesian networks using Tetrad IV which uses PC search, Bayes parametric model, ML Bayes Estimator, Bayes instantiated model, and Bayes classifier .....	133
Figure 4-33: Interactions from 31 genes and the 8 hallmarks in Train Metastasis group using Tetrad IV.....	134
Figure 4-34: Interactions from 31 genes and the 8 hallmarks in DFCI Metastasis group using Tetrad IV.....	134
Figure 4-35: Interactions from 31 genes and the 8 hallmarks in MSK Metastasis group using Tetrad IV.....	135
Figure 4-36: Interactions from 31 genes and the 8 hallmarks in Train Non-Metastasis group using Tetrad IV.....	136
Figure 4-37: Interactions from 31 genes and the 8 hallmarks in DFCI Non-Metastasis group using Tetrad IV.....	136

Figure 4-38: Interactions from 31 genes and the 8 hallmarks in MSK Non-Metastasis group using Tetrad IV .....	137
Figure 4-39: Interactions among 31 genes and the 8 hallmarks extracted from KEGG PATHWAY database and all of them are confirmed with the interactions from implication networks.....	138
Figure 4-40: Interactions among 31 genes and the 8 hallmarks extracted from NCI pathways and all of them are confirmed with the interactions from implication networks.....	139
Figure 4-41: Interactions among 31 genes and the 8 hallmarks extracted from PubMed and all of them are confirmed with the interactions from implication networks.....	140
Figure 4-42: Interactions among 31 genes and the 8 hallmarks extracted from MATISSE and all of them are confirmed with the interactions from implication networks.....	141
Figure 4-43: Evidence view of the interactions between the 31 genes and the 8 hallmarks and all of them were confirmed with the interactions from implication networks.....	142
Figure 4-44: Various sources of identification of gene interactions in STRING 8 .....	142
Figure 4-45: List of the input genes (among 31 gene signature) that were identified by STRING 8 and were displayed at the output.....	143
Figure 4-46: Network 1 generated from IPA .....	144
Figure 4-47: Network 2 generated from IPA .....	144
Figure 4-48: Network 3 generated from IPA .....	145
Figure 4-49: Network 4 generated from IPA .....	145
Figure 4-50: Network 5 generated from IPA .....	145
Figure 4-51: Merged network from all the 5 networks shown above where grey connections are the intra-network connections and orange connections are inter-network connections in IPA..	146
Figure 4-52: Interactions between the 31 genes and the 8 cancer hallmarks extracted from the merged network of all the 5 networks in IPA shown above where the yellow genes represent the major cancer hallmarks.....	147
Figure 4-53: Interactions between the 31 genes and the 8 hallmarks that were extracted from Pathway studio where each kind of line represents different kinds of relationships between the genes .....	148
Figure 4-54: Variation of number of gene interactions with threshold on weights in Training Group. The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds.....	151

- Figure 4-55: Variation of number of gene interactions with threshold on weights in DFCI test Group. The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds..... 152
- Figure 4-56: Variation of number of gene interactions with threshold on weights in MSK test Group. The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds..... 152
- Figure 5-1: Changing the directory to the current directory and compiling the C-code to generate the required dynamic linked library files to be used for executing code in R ..... 156
- Figure 5-2: C-code for the first version of the package ..... 156
- Figure 5-3: Main difference between the C-codes shown in the second version of the code..... 157
- Figure 5-4: Output from R: Red lines are the input code and the next blue lines are the outputs after execution of the entire package after around 40 minutes ..... 157

## List of Tables

Table 3-1: Number of genes identified to have interactions with major cancer hallmarks in each prognosis group in each gene signature .....	69
Table 3-2: All signatures obtained using different combinations of Hallmarks.....	71
Table 3-3: AUC's of training set (256 samples) obtained when best probes among duplicates were considered .....	73
Table 3-4: p-values from Random test of training set (256 samples) obtained when best probes among duplicates were considered .....	73
Table 3-5: AUC's of DFCI test set (82 samples) obtained when best probes among duplicates were considered .....	74
Table 3-6: p-values from Random test of DFCI test set (82 samples) obtained when best probe among duplicates were considered .....	74
Table 3-7: AUC's of MSK test set (104 samples) obtained when best probes among duplicates were considered .....	74
Table 3-8: p-values from Random test of MSK test set (104 samples) obtained when best probes among duplicates were considered .....	75
Table 3-9: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 11-gene signature .....	77
Table 3-10: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 21-gene signature .....	78
Table 3-11: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 31-gene signature .....	78
Table 4-1: Details of the 31 gene signature that have been confirmed by Dr. Yong Qian from NIOSH .....	85
Table 4-2: T-test outputs for different predictors such as Stage (Stage-2 to Stage-1 and Stage-3 to Stage-1), Tumor differentiation (Moderate to Well and Poor to Well) and Lymph node metastases (LN+ to LN-) outputs for all the 31 genes in the signature .....	88
Table 4-3: Fold changes for different predictors such as Stage (Stage-2 to Stage-1 and Stage-3 to Stage-1), Tumor differentiation (Moderate to Well and Poor to Well) and Lymph node metastases (LN+ to LN-) outputs for all the 31 genes in the signature .....	90
Table 4-4: Sensitivity, Specificity and Overall Accuracy of Training data calculated from contingency table .....	93

Table 4-5: Sensitivity, Specificity and Overall Accuracy of DFCI data calculated from contingency table .....	94
Table 4-6: Sensitivity, Specificity and Overall Accuracy of MSK data calculated from contingency table .....	95
Table 4-7: Comparison of 31 gene signature with model from Shedden et al [20] on both the Test datasets where log-rank p-values, hazard ratios, and confidence intervals were obtained from the CPE package which use the risk scores of the entire signature as input.....	95
Table 4-8: Different signatures used to compare the performance of the 31 gene signature in GSEA .....	98
Table 4-9: Different signatures Enriched in phenotype “Good”, which include the 31 gene signature.....	100
Table 4-10: Different signatures Enriched in phenotype “Poor” .....	100
Table 4-11: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the training dataset using p-values from significance test ..	103
Table 4-12: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the training dataset .....	103
Table 4-13: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the DFCI test dataset using p-values from significance test	104
Table 4-14: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the DFCI test dataset.....	105
Table 4-15: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the MSK test dataset using p-values from significance test	106
Table 4-16: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the MSK test dataset .....	107
Table 4-17: Multivariate Cox Proportional Analysis of Age, Gender, Lymph node Metastasis, Tumor size and Risk Score* .....	108
Table 4-18: Multivariate Cox Proportional Analysis of Age, Gender, Race, Smoking Status, Lymph node Metastasis, Tumor size, Tumor grade and Risk Score* .....	110

Table 4-19: Number of patients in each of the groups in each dataset along with number of censored patients.....	112
Table 4-20: Number of interactions between the 31 genes and the 8 hallmarks for various datasets in both the groups.....	113
Table 4-21: Number of differential components between both the groups for the 31 genes and the 8 hallmarks for various datasets.....	114
Table 4-22: p-values of Gene Ontology terms identified from known classes in Common interactions among Train and DFCI Metastasis group from PRODISTIN.....	119
Table 4-23: p-values of Gene Ontology terms identified from known classes in Common interactions among Train and MSK Metastasis group from PRODISTIN.....	122
Table 4-24: p-values of Gene Ontology terms identified from known classes in Common interactions among Train & DFCI Non-Metastasis group from PRODISTIN.....	125
Table 4-25: p-values of Gene Ontology terms identified from known classes in Common interactions among Train & MSK Non-Metastasis group from PRODISTIN.....	127
Table 4-26: p-values of Gene Ontology terms identified from known classes in Common interactions among DFCI & MSK Non-Metastasis group from PRODISTIN.....	129
Table 4-27: p-values of Gene Ontology terms identified from known classes in Common interactions of 3 datasets Metastasis group from PRODISTIN.....	130
Table 4-28: p-values of Gene Ontology terms identified from known classes in Common interactions of 3 datasets Non-Metastasis group from PRODISTIN.....	132
Table 4-29: Comparison of number of interactions from Poor and Good Prognosis of each dataset generated in Implication Networks and Bayesian Networks (using Tetrad IV).....	149
Table 4-30: Comparison of number of interactions among the 31 genes and the 8 hallmarks identified from different biomedical tools found in implication networks and Bayesian networks.....	149
Table 4-31: Number of Biological Processes identified using Prodistin when interactions from implication networks and Bayesian networks are given as input.....	150
Table 4-32: Comparison of number of interactions from Poor and Good Prognosis of each dataset in Implication Networks and Bayesian Networks (using Tetrad IV) with application of thresholds on weights.....	150



## 1 Introduction

Lung Cancer is caused due to the uncontrolled growth of cells in the tissues of lungs. It is critical to identify gene signatures that can predict cancer recurrence to improve patient care. Genes having high degree of connections with the major cancer markers have strong impact on the network topology [7] and are thus the critical genes of the network. There are different techniques which can identify these critical genes.

Feature selection techniques have been used earlier to find prognostic markers from a group of data by eliminating genes which have little or almost no predictive information [47]. These techniques were used in machine learning particularly for the purpose of removing irrelevant or redundant features from data and forming a subset of relevant features. Though feature selection techniques have a good number of advantages, they still have a few limitations. When there are a large number of features, the search for a good subset of features (which provides optimal results) becomes very complicated and tedious. Moreover feature selection techniques consider the behavior of genes individually which might not act in the same manner in the presence or absence of other genes.

Network-based techniques can be used to find gene signatures and overcome the limitations of feature selection methods. Network-based techniques work in uncovering the causal relationships between the genes and are also better in terms of stability and classification accuracy [7] and thus they are an efficient means of finding prognostic gene signatures when compared to feature selection algorithms. They consider the signature of genes as a whole instead of considering each gene individually and thus emphasize on the molecular interactions

between pairs of genes. This works well as genes might not act in the same way when they are alone and when they are acting along with other regulators. Network-based techniques are more useful in cases where huge datasets come into picture [8]. This is due to the fact that performing an exhaustive or complete feature selection technique on a huge dataset would be very time taking and also requires a lot of resources. Most of the lung cancer datasets are huge and thus using these network techniques helps in identifying signatures faster and in an accurate manner and it also helps in analyzing the signatures in a better way.

Currently, there are different network-based techniques that are in use such as coexpression network, Bayesian network, and artificial neural network. Though these network-based techniques overcome the limitations of feature selection methods, they still have a few limitations each. Correlation-based coexpression networks are inconsistent as their accuracy decreases with increase in network size [7]. Bayesian networks cannot model feedback loops and their complexity increases exponentially with the number of genes in the network [4]. Artificial neural networks are very complex and time taking in nature. Moreover to our knowledge, neural networks have not been used for modeling molecular interactions yet.

To overcome the limitations of the currently used network-based techniques, implication networks based on prediction logic were chosen to generate the genome wide networks [2]. The methodology used in implication networks is computationally manageable for analyzing large datasets and integrates formal logic and statistics [1], thus making it more efficient.

To generate the genome wide networks based on prediction logic [2], the gene expression data (from University of Michigan Cancer Center (UM) and Moffitt Cancer Center (HLM) together used as training dataset [20]) was separated in to two groups (metastasis: corresponds to the high risk

group and non-metastasis: corresponds to the low risk group) based on the survival period and survival status. The genome wide networks of both the groups were compared and the common interactions they have were removed. Thus we could focus on the differential components in networks that remained which are the interactions that differentiated the metastasis group from the non-metastasis group.

To identify prognostic signatures, major cancer hallmarks such as E2F, EGF, EGFR, KRAS, MET, RB1, and TP53 were considered and the genes that interact with all these major cancer hallmarks were considered to form a signature. The hallmark E2F had many probes like E2F1, E2F2, E2F3, E2F4, and E2F5. These probes can be picked in various combinations depending on their functionality. Thus different sets of hallmarks can be considered. Different signatures can be identified by varying the set of hallmarks used to pick the genes. Thus different gene signatures were identified based on the interactions between the genes and the hallmarks under diseased conditions.

To identify the most prognostic signature from the obtained signatures, survival analysis [35] was done using techniques such as Time-dependent ROC [16] (statistical p-values and area under curves (AUC) over time were used as measures to compare the different signatures), Random testing (different signatures as the same size of the identified signatures were picked randomly and checked where our signature stands among the randomly picked signatures), and COX proportional hazard model [18, 19, 28] (Kaplan-Meier plots and log-rank test results were observed). For analysis with COX proportional hazard model, both univariate (considering the gene expression values of the genes only) and multivariate analysis (considering the risk scores of the signatures as a predictor and comparing with other predictors such as age, gender, smoking status, tumor size etc. with and without the risk scores) were done. Kaplan Meier plots

were used for determining the significance of the signature in differentiating the two groups from one another. Log-rank p-values were observed from the COX proportional hazard model and signature was picked which had values less than 0.05 for training and test sets, showing it to be significant. Multivariate analysis using the COX proportional hazard model was done as a part of the evaluation of the signature with respect to other clinical parameters.

To validate the signature obtained, both prognostic and topological validation was performed on the test datasets (from Memorial Sloan-Kettering Cancer Center (MSK) and the Dana-Farber Cancer Institute (CAN/DF) [20]). The prognostic validation was conducted using techniques such as Overall Accuracy [32], Concordance Probability Estimate (CPE) [29], and Gene Set Enrichment Analysis (GSEA) [30]. Sensitivity and specificity were measured along with the overall accuracy values. CPE was used to evaluate the distinguishing power and the predictive accuracy of the statistical model. CPE measures included the statistical log-rank p-values, hazard ratios, and 95% confidence intervals which were compared with published signature from Shedden et al [20]. The results showed that the 31-gene signature had more significant statistical p-values, higher hazard ratios, and higher CPE values which confirm that the signature is better when compared to the other published signature. GSEA is a powerful analytical method that computed whether the 31 gene signature is statistically significant and whether the gene set has agreeable differences between the two phenotypes (biological states). GSEA was used to compare our signature with many other previous signatures using False Discovery Rates (FDR) and Normalized Enrichment Scores (NES). GSEA results showed that the signature had  $FDR < 0.25$  which makes it significant. The comparisons above showed that our signature was either comparable or better than the other signatures on the same datasets.

To topologically validate the gene signature, the interactions from implication network were compared with interactions from Bayesian network generated by Tetrad IV<sup>1</sup>. Then various tools such as Prodistin<sup>2</sup>, KEGG<sup>3</sup>, NCI<sup>4</sup> pathways, PubMed<sup>5</sup>, Matisse<sup>6</sup>, String<sup>7</sup>, Ingenuity Pathway Analysis<sup>8</sup>, and Pathway studio<sup>9</sup> were used. These tools extracted their interactions from various sources such as literature, curated databases, etc. All the interactions found from the above mentioned tools were compared with the interactions generated from implication network and interactions from Tetrad IV (Bayesian network).

From the interactions extracted from various biomedical tools, it was concluded that implication networks are capable of generating many more gene or protein interactions which were validated by the molecular interactions from other tools when compared to the Bayesian networks. The functional classes identified from the signature reveal that the genes are not just structurally connected but also have biological relationships. Thus these genes could be focused in predicting cancer recurrence in therapeutic conditions.

The chapters in this thesis are as divided as follows. The second chapter provides literature review of the currently used techniques. It also provides descriptions of all the methods and web-based tools used in this study. The third chapter describes the methodology used to identify the gene signature from the genome wide coexpression networks. The fourth chapter discusses the results obtained from prognostic and topological validation techniques. The fifth chapter describes the implementation of the software used to generate the results in both C and R<sup>10</sup>. It

---

1. <http://www.phil.cmu.edu/projects/tetrad/>
2. <http://crfb.univ-mrs.fr/webdistin/>
3. <http://www.genome.jp/kegg/>
4. <http://pid.nci.nih.gov/>
5. <http://www.ncbi.nlm.nih.gov/pubmed/>
6. <http://acgt.cs.tau.ac.il/matisse/>
7. <http://string.embl.de/>
8. <http://www.ingenuity.com/>
9. <http://www.ariadnegenomics.com/products/pathway-studio/>
10. <http://www.r-project.org>

also describes the versions of the editors and the configuration of the system used to run the analyses. The sixth chapter concludes all the above mentioned chapters and also includes the prospective work that will be carried out relating this approach.

## **2 Background**

### **2.1 Introduction**

This chapter describes various techniques to identify signatures. Feature selection methods are described in brief followed by their limitations. These limitations are overcome by introduced network-based techniques. Different network-based techniques currently used such as correlation-based coexpression networks, Bayesian belief networks, and artificial neural networks are discussed followed by their limitations which are overcome by the implication networks. The implication networks are discussed and the algorithm which has been used to induce the implication networks is discussed. Different validation techniques which have been used to validate the signature found from implication networks were discussed. Finally a summary of the entire chapter is given.

### **2.2 Different techniques to identify signatures**

There are different procedures to identify gene signatures. Potential markers have been screened earlier by identifying the overexpressed or the underexpressed genes. But this process is not good enough as the information of each individual gene is considered when the interactions between genes were supposed to be considered [15].

Feature selection [47] techniques have been used earlier to find prognostic markers from a group of data by eliminating genes which have little or almost no predictive information. These techniques were used in machine learning particularly for the purpose of removing irrelevant or redundant features from data and forming a subset of relevant features. They help in overcoming

the curse of dimensionality by reducing the number of features that have to be considered and thus speeding up the process. They can be used with both supervised (to produce high classification accuracy) and unsupervised learning (to find good subsets of features that form quality clusters). Though feature selection techniques have several advantages such as removal of redundant and irrelevant features, improving the classifier performance, etc., they still have a few limitations. When there are a large number of features in the beginning, as in the case of lung cancer genes, feature selection techniques become very complicated and it becomes tedious to find a good subset of features. Moreover they consider the genes individually which might not act in the same manner in the presence or absence of other genes.

The most important advantage of network-based approaches over feature selection methods is that they can capture and represent more complex types of relationships among genes or any variables of interest [8]. Since there will be a large number of relationships between genes, methods other than network-based procedures become more complex and the computation of such models becomes very tedious (for example, in case of feature selection methods, optimal output requires exhaustive search which is very time consuming). Network-based techniques help in revealing the underlying molecular mechanisms related to the genes. Networks built with genes can be used to identify disease mechanisms [34] and for drug discovery [33], and also for identifying prognostic subnetworks which lead to metabolic pathways [33]. Other methods (such as feature selection techniques) ignore genes which do not have significant differential expression individually in different classes but which actually play vital roles as a member of a group in certain pathways.



Gene networks are constructed in such a way that any pair of genes are connected if some measure (calculated from the current conditions of the genes) related to both the genes exceeds a given threshold [13].

There are many network-based approaches that are already in use for classification analyses and for identifying interactions between genes. Some of them are described below.

### 2.2.1 Coexpression Network

Gene coexpression network connects genes with similar expression profiles (such as the Pearson correlation coefficient [15] or the clustering coefficient [13]) and thus connects functionally related genes [13, 15]. This network tries to investigate the transcriptional changes in terms of “gene interactions” rather than at the level of “individual genes”.

Pearson correlation coefficient measures the degree of linear dependence between two time-courses of gene expression levels [14]. It is close to one when there is good correlation between the time series. It is near negative one when there is negative correlation and is close to zero when there is no correlation between the expression values [42]. It can be calculated as shown below. If  $\rho$  stands for correlation between the genes  $g_i$  and  $g_j$ ;  $G_{ik}$  and  $G_{jk}$  are the gene expression values of genes  $g_i$  and  $g_j$  respectively;  $\mu_i$  and  $\mu_j$  are the means and  $\sigma_i$  and  $\sigma_j$  are the standard deviations of the genes  $g_i$  and  $g_j$  respectively;  $N$  is the total number of genes; then Pearson correlation ( $r$ ) is defined as

$$r = \rho(g_i, g_j) = \frac{1}{N} \sum_k \left( \frac{G_{ik} - \mu_i}{\sigma_i} \right) \left( \frac{G_{jk} - \mu_j}{\sigma_j} \right) \quad (2.1)$$

Hence Pearson correlation ( $r$ ) was calculated for each dataset and was converted in to a standard normal metric using the Fisher's transformation [15]. This standard normal metric shown below is called effect size ( $z$ ) which was used as a measure of treatment or covariate effect.

$$z = 0.5 \log \frac{(1+r)}{(1-r)} \quad (2.2)$$

Clustering is generally used to cluster (group) genes based on a correlation-based distance measure quantifying the degree of co-regulation [14]. Thus the function of an unknown gene can be predicted from the known functions of other genes present in the same cluster [14]. Clustering algorithms work well when the genes are co-regulated. Gene expression clusters can also be mapped on to metabolic networks in order to discover pathways of interest. The clustering coefficient of gene  $i$  is denoted by  $C_i$  and is calculated as shown below. If  $k_i$  is the number of first neighbors of gene  $i$  and  $E_i$  is the number of edges between the  $k_i$  first neighbors, clustering coefficient of the entire network can be calculated by taking the average of the clustering coefficients of all the genes in the network as shown below.

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \quad (2.3)$$

The underlying postulation of the network distance metric is that the enzymes are related according to their proximity in the network. If this metric is above a specific threshold, the pair of genes would be connected. It considers that a rise or fall in the correlation of a gene pair might be associated with the upregulation or downregulation of other genes in the same functional cluster. These networks constructed from pair-wise correlation coefficients have provided a productive procedure to recognize functional transcriptional modules related with specific biological processes [6].

Aoki et al. [6] explored the gene co-expression networks in plant biology and concluded that co-expression network analysis provided innovative awareness in the system level understanding of plant biology. In many cases, gene co-expression networks implied the presence of functional linkage between genes associated with biological processes.

L.L. Elo et al [13] proposed a systematic approach for the estimation of the threshold of coexpression networks directly from their topological properties. They used the clustering coefficient for the threshold selection which when gradually increased reduced the number of links from the initially complete graph of coexpression networks. They experimented on the simulated data generated using the stochastic model of Thalamuthu et al [49] which consisted of 60 datasets. The biological relevance of the coexpression was investigated by the p-values.

Hanisch et al. [14] proposed the construction of a distance function (correlation-based distance function) which combined the information from biological networks (in an integrated manner) and gene expression data. They focused on the analysis of co-regulated metabolic pathways which were supported by gene expression measurements. They calculated the Pearson correlation coefficient on log-ratio transformed data and then converted it in to a distance metric which quantified the degree of dissimilarity of their gene expression dataset. They defined a graph distance function on the networks and combined it with correlation based distance function for gene expression measurements. They conducted the experiments on the organism *S.cerevisiae* (yeast).

Choi et al [15] introduced a model (mentioned above using Pearson correlation coefficient and its Fisher transformation to find effect size) for finding the differential coexpression from microarrays and testing its biological validity with respect to cancer. They collected data from 10

published gene expression datasets from cancers of 13 various tissues and built 2 different coexpression networks, a tumor network and a normal network which were compared.

S. Tornow and H. Mewes [41] proposed a technique which was based on collective, multi-body correlations in a genetic network. They calculated the correlation strength of a group of genes in a coexpression network which were identified as members of a module in another protein interaction network and estimated its correlation probability.

Zhang and Horvath [43] proposed a general framework for soft thresholding which assigned a connection weight to each gene pair. They used several adjacency functions (such as sigmoid function, power adjacency function, etc.) to convert the correlation coefficients to connection weights. They experimented on simulated data, a cancer microarray dataset and a yeast microarray dataset.

Thus coexpression networks have been used in several applications such as for discovery of genetic modules, applying to human T helper cell differentiation process [13], for topology based cancer classification [7], for molecular characterization of cellular state, etc.

There are a few limitations of coexpression networks. Correlation-based coexpression networks are based on similarities and clustering based coexpression networks are based on the distance measures. Thus they do not provide a predictive structure and do not infer causal relationships among genes. High correlation is exhibited by genes when the entire set of expression patterns across different conditions is similar. On the other hand, high correlation is also exhibited by genes if they are expressed together under a few conditions and are otherwise silent [6]. Moreover the accuracy of correlation-based coexpression networks decreases, as the network size increases and it is highly inconsistent.

### 2.2.2 Bayesian Network

The Bayesian networks model is a causal network which represents the joint probability distributions. Bayesian networks are useful for describing complex probabilistic models which require learning from noisy observations. Bayesian Networks are thus capable of estimating the confidence in different features of the network and thus are a promising tool for examining gene expression patterns [4].

If we have a finite set of random variables,  $X = \{X_1, \dots, X_n\}$ , where  $X_i$  is a variable which might take values from the domain  $\text{Val}(X_i)$ , Bayesian networks are represented using joint probability distributions consisting of two components,  $B = \langle G, \theta \rangle$ ; a directed acyclic graph [4] (DAG)  $G$  (whose vertices correspond to the random variables,  $X$ ) and a conditional distribution for each variable  $\theta$  (given its parents in  $G$ ). According to Markov assumption, each variable  $X_i$  is independent of its non-descendants, given its parents in  $G$  and their joint probability distribution can be defined as below [4].

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \mathbf{Pa}^G(X_i)) \quad (2.4)$$

Here  $\mathbf{Pa}^G(X_i)$  is the set of parents of  $X_i$  in  $G$ . Once networks are built, they are needed to be scored by some means so that the networks are evaluated and the optimal network can be found. Posterior probability can be used to evaluate the graphs. If a large number of networks are given, learning procedures can pinpoint the exact network structure which has best dependencies in the distribution.

Bayesian Networks were used to describe the interactions between genes in a paper by Friedman et al. [4] where they described a method to recover the gene interactions from microarray data.

They also applied the method to the *S.cerevisiae* cell-cycle measurements of Spellman et al. (1998). They used priors described by Heckerman and Geiger (1995) for hybrid networks of multinomial distributions and conditional Gaussian distributions.

It has been described by Friedman et al. [4] that a causal network models the distribution of the observations as well as the effects of interventions.  $X \rightarrow Y$  and  $X \leftarrow Y$  are equivalent in Bayesian networks but they are not equivalent in causal networks. If  $X$  causes  $Y$ , then changing the value of  $X$  affects the value of  $Y$ . But it is not true the other side, i.e., changing the value of  $Y$  does not affect the value of  $X$ . Their approach was to analyze a high number of high scoring networks which requires an efficient learning algorithm such as the Sparse Candidate algorithm. To relate their analysis with the biological phenomena in the data, they used the order relations and Markov relations found from their data.

As Bayesian networks have the capability of working even in highly noisy surroundings, it has many real-world applications. Some of them in bioinformatics are for building gene regulatory networks and protein structures [34]. They are also applicable to other fields such as medicine, image processing, information retrieval, etc.

There are a few limitations to the Bayesian network approach. Since the Bayesian networks are directed acyclic graphs, the probabilities of the child nodes are calculated from the parent nodes. Thus Bayesian networks cannot have loops and they also require a subjective prior (for the first parent node). Bayesian networks need complete knowledge of the real-world in order to build the correct causal model. These networks are expensive to compute and the rate of complexity increases exponentially with the number of genes present in the network [8]. Thus they become more impractical and inappropriate.

TETRAD IV<sup>11</sup> and its search algorithms were developed with the support of National Aeronautics and Space Administration and the Office of Naval Research. TETRAD IV is a program for working on causal/statistical models particularly Bayesian belief networks. It is used for creating, simulating data from, estimating, testing, predicting with and searching for causal models. It has a friendly interface and no programming knowledge is required to use it. It is unique in the suite of principled search algorithms.

A program description of a causal model is done in three stages in TETRAD IV. The first one is a picture which uses a directed graph to state in detail the hypothetical causal relations among variables. The second stage would be to specify the family of probability distributions and the kinds of parameters associated with the graphical model. The final stage would be to specify the numerical values of the parameters explained earlier.

Sessions in TETRAD IV are built by dragging boxes in to the workspace and then connecting them with arrows in legal ways that represent their dependencies. The Figure 2-1 below shows the model used in TETRAD IV to build Bayesian networks. This network was compared to the implication network built using prediction logic.

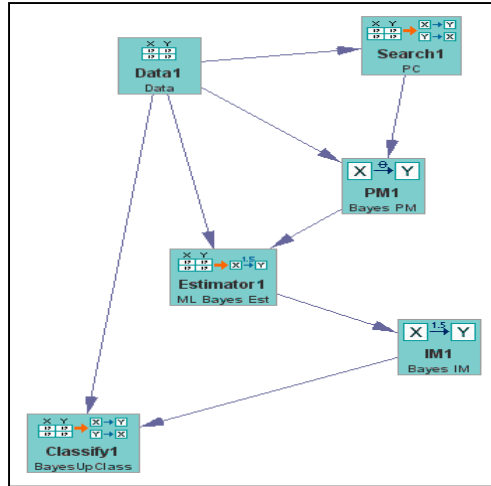


Figure 2-1: Model used in TETRAD IV showing all the boxes

**Data box:**

We use data that was loaded from an external file. Here “Data set” list is a record of available datasets, where one of the lists is considered as “selected”. There are three types of data that can be stored in the data set list namely: Tabular data set, Covariance matrix, and Correlation matrix. We use Tabular data set. The Data wrapper is shown in Figure 2-2 below.

Data1 (Data Wrapper)										
File Edit Knowledge Tools										
5YR18*13.csv										
	C1-T	C2-T	C3-T	C4-T	C5-T	C6-T	C7-T	C8-T	C9-T	
	MULT	215642_at	217470_at	ACTL6B	ACTR8	ADAM3B	BCDN3	C1orf68	CAP2	CDKN2B
1	1	0	1	0	0	1	0	1	1	0
2	1	0	1	0	1	0	1	1	0	1
3	1	0	0	0	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0	0
5	1	0	0	0	0	1	0	1	1	1
6	1	0	1	0	0	1	1	1	0	1
7	1	0	1	0	1	1	0	1	0	0
8	1	0	0	0	0	1	1	1	0	1
9	1	0	1	0	0	0	1	0	1	0
10	1	0	1	0	0	1	1	1	0	1
11	1	1	0	0	0	0	1	1	0	1
12	1	0	0	0	0	0	1	1	0	0
13	1	0	1	0	0	0	0	0	0	1
14	1	0	0	0	0	0	0	1	0	0
15	1	0	1	0	1	0	1	1	0	1
16	1	0	1	0	1	1	1	0	0	0
17	1	0	1	0	1	1	0	0	0	1
18	1	0	1	0	0	0	1	1	0	1
19	1	0	1	0	1	0	1	1	1	0
20	1	0	1	0	0	0	0	0	1	1

Figure 2-2: Data Wrapper shown



### Search Box:

TETRAD IV has a variety of search algorithms to assist in searching for causal explanations of a body of data. The search algorithms read in data and return information about a collection of alternative causal graphs that can explain features of the data. Some search algorithms can often predict whether a particular variable influences another or not. Search algorithms do not output an estimated model with parameter values; instead they output a description of a class of causal graphs that explain statistical features of the data which were considered by the search procedures. Some of the search procedures available are PC, CPC, PCD, FCI, etc which are shown in Figure 2-3 below. We use PC technique which searches for Bayes net or SEM models when it is assumed there is no unrecorded variable that contributes to the association of two or more measured variables. The output obtained after execution of PC search algorithm is shown in the Figure 2-4 below.

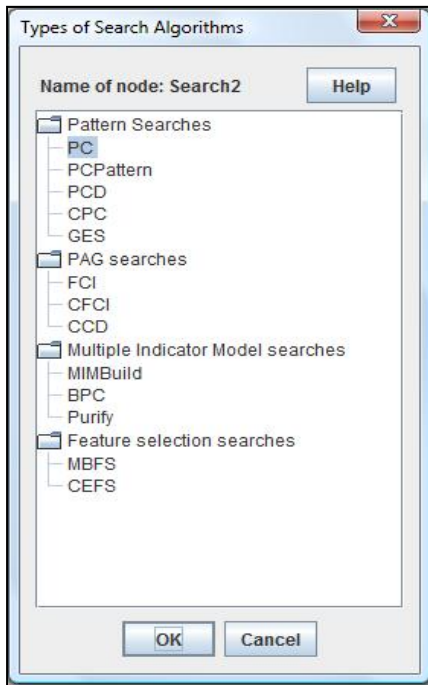


Figure 2-3: Available Search Algorithms

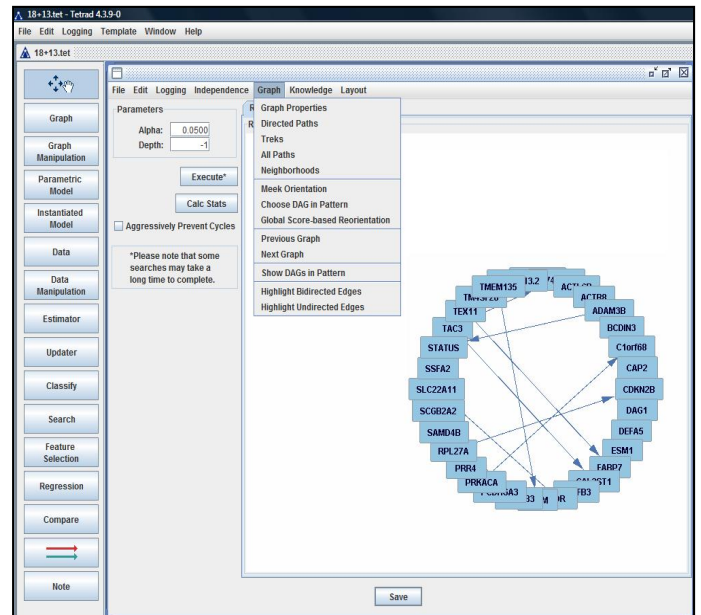
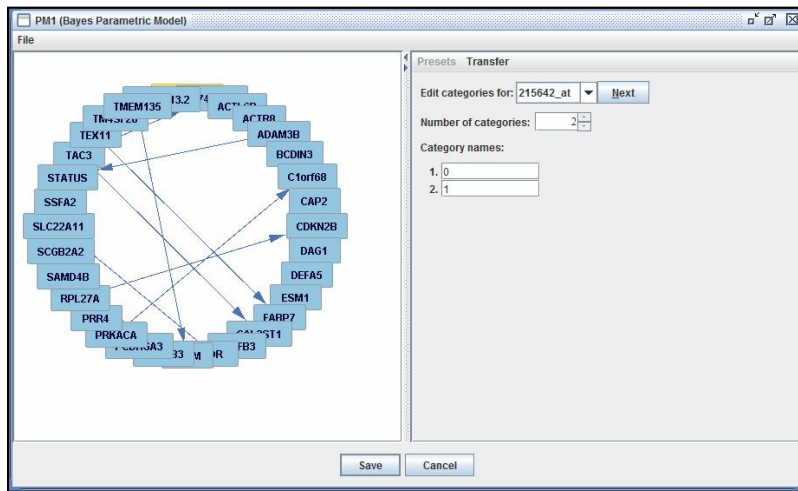


Figure 2-4: PC Search after execution

***Parametric Model Box:***

A parametric model specifies the family of probability functions connecting cause and effect, but does not specify values for its parameters. Two types of parametric models can be created using TETRAD IV namely Bayes and SEM. If Bayes net is chosen, then the input graph to the PM box will be parameterized as a categorical model in which the parameters are the unspecified conditional probabilities of values of each variable on the values of its parent variables in the graph. Bayes PM takes a DAG and adds to it, two bits of information (the number of categories and the list of categories). If SEM is chosen, then the graph will be parameterized as a linear Gaussian model with variances and linear coefficients. The Bayes PM is shown in the Figure 2-5 below.

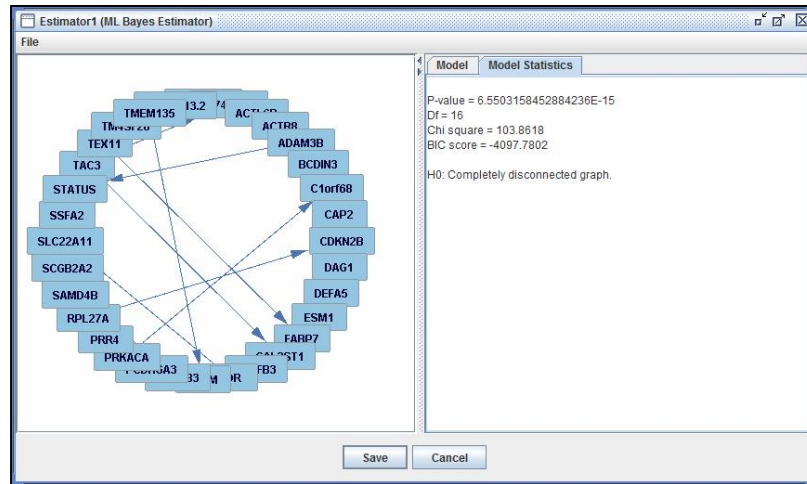


**Figure 2-5: Bayes Parametric Model**

***Estimator Box:***

The Estimator box takes in information from the Parametric Model and the Data and outputs an instantiated model. The procedures in the statistical estimator allow estimation of the parameters based on the input data. Types of estimators include ML estimator, SEM estimator and Dirichlet

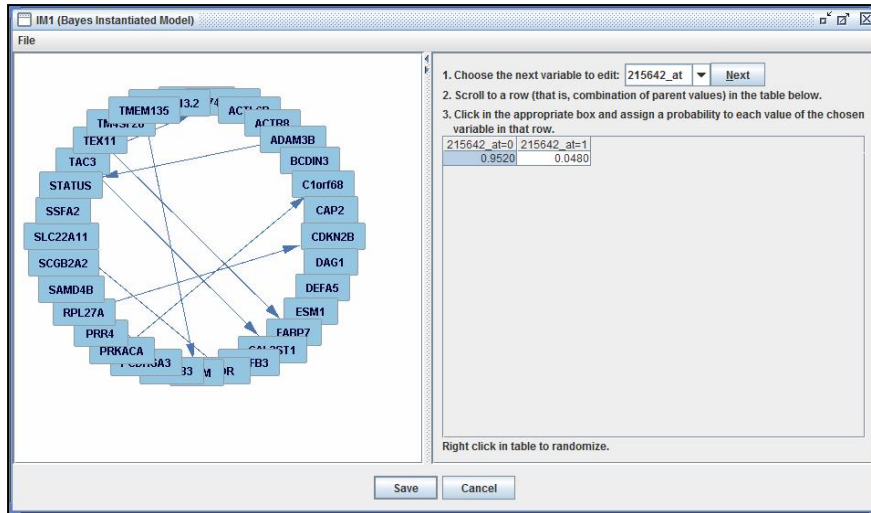
estimator. There are also procedures for handling missing values in the input data. The ML Bayes estimator is shown in the Figure 2-6 below.



**Figure 2-6: ML Bayes Estimator**

***Instantiated Model Box:***

An Instantiated model specifies particular numerical values for the parameters of a parametric model. There might be three types of instantiated models namely Bayes instantiated model, Dirichlet Bayes instantiated model, and SEM instantiated model. A Bayes instantiated model extends a Bayes parametric model, specifying values of the parameters in the Bayes net. The parameters for a Bayes net are the conditional probabilities stored in the conditional probability tables, one for each variable in the Bayes net. The Bayes instantiated model is shown in the Figure 2-7 below.



**Figure 2-7: Bayes Instantiated Model**

***Classify Box:***

A Classifier box requires input from the Data and from IM box. It is used to classify new cases with the Bayes net in the IM box. The user specifies a target variable in the IM and the classifier uses the Bayes net structure of the IM to predict the values of the target in the data set. Statistics on the classification accuracy are provided as ROC curves (shown in Figure 2-8), AUC and confusion matrices (shown in Figure 2-9).

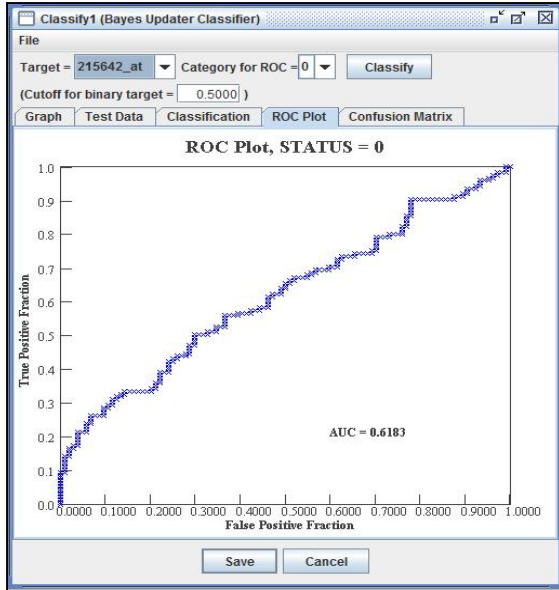


Figure 2-8: ROC curve after classification

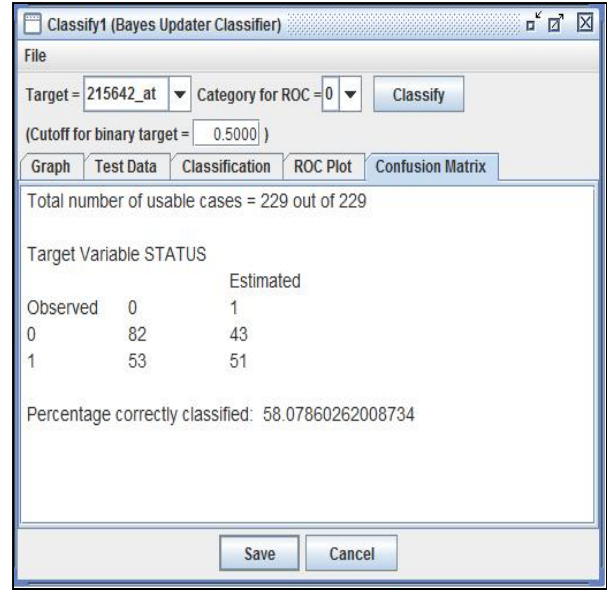
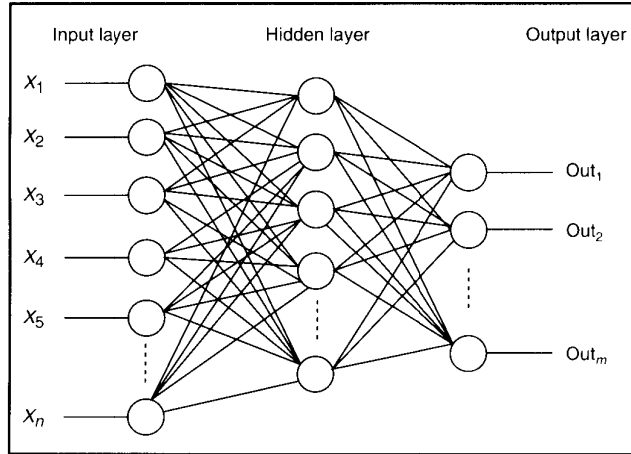


Figure 2-9: Confusion Matrix after Classification

### 2.2.3 Artificial Neural Network

Artificial neural networks are computing systems which try to mimic the elements and structures of the nervous system in a summarized manner [50]. They were actually developed as a better means of understanding the human brain and then they were used for roles like optimization [9]. In other words, if a neural network is given a large set of information, it can generalize from that data by learning about it (training). This network is built on the strategy of train, test, differentiate, and retrain on reduced gene set and then retest [10, 12].

An artificial neural network is a group of nodes and lines between the nodes where each node depicts a neuron and the lines depict the relationship between the neurons. Strength of each relationship is defined by a variable on which threshold can be applied to remove insignificant relationships. These nodes and their interconnections are organized as layers. There will be an input layer (to which the input is presented), a hidden layer (where all the processing is done on the incoming data), and an output layer (where output is retrieved) as show in figure below.



**Figure 2-10: Structural diagram of a general artificial neural network**

The learning stage (training) of the neural networks uses a dataset which has both the input and the output. When the input is fed in to the network and an output is found, it is compared with the already present output for errors (test and differentiate) and then these outputs are fed as inputs, again and again, till the minimum error requirements is reached (retrain) as shown in Figure 2-11. This is also called back propagation as results obtained once are re-fed to the inputs.

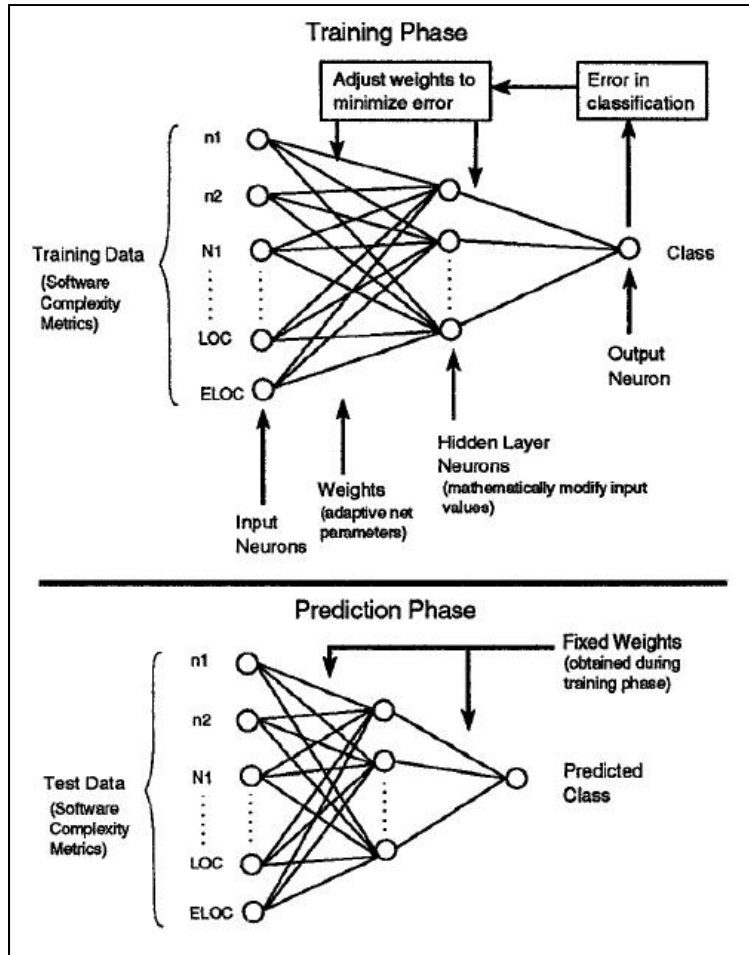


Figure 2-11: Training and Prediction Phases in a model built with Artificial Neural Networks [10]

Neural networks are particularly useful for classification analyses that are highly tolerant to precision errors [11]. They can be used as alternatives to approaches which are limited by assumptions of normality and linearity.

Goodman and Harrell [9] discussed the advantages and limitations of using neural networks for biostatistical modeling. They compared the neural network model with the generalized linear model which is another popular biostatistical method. They found out that for binary outcomes such as survival, cancer recurrence, etc. a link function is required to monotonically constraint

the output prediction. The neural networks had efficient dimensional scaling but they had an increased computational burden to optimize the model.

Boger [44] demonstrated the application of artificial neural networks for gene array analysis and cancer cell identification. The data was first trained using a Principal Component Analysis training algorithm and then local minima avoidance and escape algorithms were used. The inputs were then ranked according to their relevance to the artificial neural network prediction accuracy and the least relevant inputs were discarded. The remaining set of inputs was retrained to get better prediction accuracy and this process is repeated.

Xu et al [45] discussed the method of distinguishing between two kinds of cancers using artificial neural networks and gene filtering. In this method, the data was first clustered and it was filtered using SAM gene filtering. The artificial neural network was then constructed based on the principle of FeedForward with Error Backpropagation.

Keedwell et al [46] discovered a neural-genetic method which combines a genetic approach with a supervised single layer artificial neural network to form a hybrid system. In this approach, they formed a training set on which the gradient descent algorithm was applied via the artificial neural network to determine the weights between the input genes and the output genes. The output is tested for errors and the process is repeated until the errors meet the stopping criterion. They experimented on the yeast *S.cerevisiae* data which consisted of 2468 genes.

Neural networks have been widely spread in various fields [9] such as pattern recognition, speech synthesis, robotic control, etc. They can also be used to identify most relevant genes from gene expression data, also to identify the high risk program modules [10] in software engineering applications.



Neural networks have a few limitations. Neural networks need a very huge training dataset to generalize. If the dataset is not sufficiently large enough, the network model will be biased. They cannot be used for data which do not have any correlation among the variables present in the data. Moreover these networks need a lot of iterations to reach an approximation with minimum errors. Since the iterative process is a time taking procedure, the amount of time required for different networks is not always the same and hence it is a major shortcoming of neural networks. Sometimes the neural networks might be over-trained which might lead to good results only in the training but which actually don't work for the test datasets. Neural networks are difficult to understand and are not easily extensible. Neural networks are considered to be black boxes [44] as the process that is going on in the hidden layer is not known to the user. Moreover many applications of artificial neural networks include classification analyses but to our knowledge, there are no applications for the complete modeling of gene-gene interactions yet.

#### **2.2.4 Boolean Networks**

Boolean networks are a kind of dynamic networks which are used to model gene regulatory networks.

Sahoo et al [3] proposed a method for extracting the Boolean implications from large microarray data. They analyzed the data from three species: humans, mice and fruit flies. They tried to capture new relationships that were preserved in all the three species in spite of the differences in various factors like tissue difference, gender differences, etc.

Boolean networks are limited to small scale networks. Since they are dynamic networks, it becomes very difficult to model them at genome scale. This is due to the exponential increase in computation with the number of entities [34].

## 2.3 Implication Networks

Since all the above mentioned network-based approaches have a few limitations, another type of network called the implication network is considered to build the interactions between the genes. The algorithm to induce an implication network was first developed by Liu et al [1]. This algorithm was based on binomial distribution. An alternative algorithm to induce the network (which can be used not just with binomially distributed data as mentioned by Liu et al [1], but in general to all implication networks built on either binomial or non-binomial data) was developed by Guo et al [2] which was based on prediction logic.

Liu et al [1] described an algorithmic means for inducing implication networks from empirical data samples. Several Monte-Carlo simulations were conducted to examine the effectiveness and validity of the induction method. Dempster-Shafer belief updating scheme was used to predict the values in implication networks.

Guo et al [2] proposed a novel methodology for predicting fault prone modules by using Dempster-Shafer methodology. This methodology was applied on two case studies based on NASA datasets and the performance of the methodology over other analyses was observed. The prediction logic induced network in this paper has been used to build the implication networks for our study.

In spite of the existence of many other network-based techniques, Implication networks were used for this study. This is because they overcome the limitations of various other network-based techniques. Implication networks are better than correlation networks in the sense that most of the interactions between the genes in implication networks have comparable correlation coefficients [3]. Thus it can be concluded that gene pairs with high correlation coefficients are

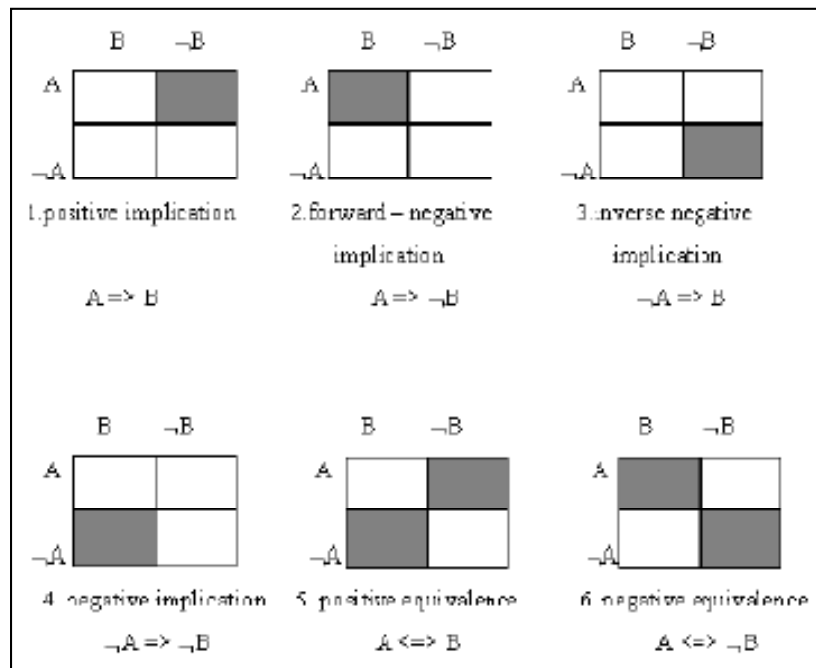
almost always present in implication networks [3]. In addition, the accuracy of the correlation networks decreases with increase in the size of the network and they cannot tolerate noise accumulations. Thus gene networks with implication relationship are superior to those with correlation relationship in terms of accuracy and stability of the classifier performance [7]. Moreover computing an implication network is not time taking as in the case of the neural networks. They are simple to construct and are fast in terms of computation. The implication networks do not need any prior learning regarding the implications, and their complexity does not increase in an exponential manner and thus they overcome the limitations of Bayesian networks.

A graph which involves nodes and arcs connecting each of the nodes in a directed manner is called an implication network. In this network, each node represents a variable which might be a gene or protein. Each arc between the nodes indicates the presence of a relation (a direct implication like influence, binding, regulation, etc.) between the nodes (genes or proteins) it connects. These arcs relate the values of each node with its parent nodes and child nodes and these values are updated at regular intervals. The arcs are accompanied by weights which represent the strength of the node relationships.

Contingency tables [2] are a tabular representation of categorical data which are used to record and analyze the relationship between two or more variables. Thus it represents the strength of association among the variables. In our network we used the contingency table to represent the occurrences of errors in samples that are associated with the six possible implications.

An implication can be defined in the following manner [1, 2]. For  $A \Rightarrow B$ , If  $A$  is True, then  $B$  is also True. If  $A$  is False, then  $B$  can be either True or False. So the erroneous case for  $A \Rightarrow B$

would be A being True and B being False. This is shown below as positive implication in the Figure 2-12 with a shaded cell. Similarly for  $A \Rightarrow \neg B$ , the erroneous case would be both A and B being True, which is shown as forward negative implication in Figure 2-12. Similarly inverse negative implication and negative implication can be understood. For  $A \Leftrightarrow B$ , A and B should both be True or both be False. So it combines the positive implication and the negative implication to form the positive equivalence. Similarly for  $A \Leftrightarrow \neg B$ , A and B should be opposite to one another. This combines the forward negative implication and the inverse negative implication to form the negative equivalence. Thus all the six relation types can be explained.



**Figure 2-12: Six possible implications relating two variables [2]**

The Contingency table that was used to calculate different values is shown below in Figure 2-13. Each of the cells represents the errors that occurred while finding the implication between the two variables.

	$B$	$\neg B$
$A$	$N_{A \rightarrow B}$	$N_{A \rightarrow \neg B}$
$\neg A$	$N_{\neg A \rightarrow B}$	$N_{\neg A \rightarrow \neg B}$

**Figure 2-13: Contingency Table**

The implication induction algorithm is shown in Figure 2-14 which was taken from Guo et al. First significance level  $\nabla_{min}$  and a minimal  $U_{min}$  are set for each  $node_i, i \in [0, n_{max} - 1]$  and  $node_j, j \in [i + 1, n_{max}]$ . Here  $n_{max}$  is the total number of attributes. Contingency table is computed for all the possible sample cases and  $Max U_p$  that satisfy the condition on  $U_{min}$  is computed for all relation types. This process is iterated till a solution exists and once a solution is found the value is returned.

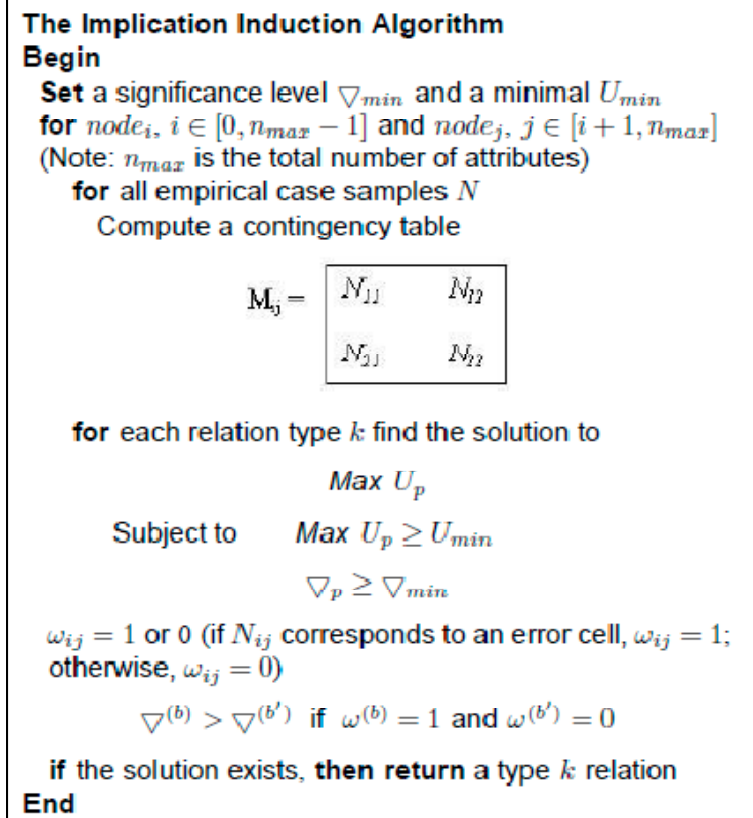


Figure 2-14: Implication induction algorithm from Guo et al [2]

In this algorithm,  $U_{min}$  and  $\nabla_{min}$  correspond to  $U_{min}$  and  $del_{min}$  respectively in our project. They are the minimum scope and minimum precision that are required for the implication rule to be considered as significant. They are calculated from simple Z-test for a cutoff value of = 1.64 . These values keep varying with the number of samples in the group.

All the values of  $U_p$  and  $del_p$  are calculated as shown below [2].  $N$  is the total number of samples.

$$U_1 = \frac{(N_{A \rightarrow B} + N_{A \rightarrow -B}) * (N_{A \rightarrow -B} + N_{-A \rightarrow -B})}{N * N} \quad (2.5)$$

$$U_2 = \frac{(N_{A \rightarrow B} + N_{A \rightarrow -B}) * (N_{A \rightarrow B} + N_{-A \rightarrow B})}{N * N} \quad (2.6)$$

$$U_3 = \frac{(N_{-A \rightarrow -B} + N_{A \rightarrow -B}) * (N_{-A \rightarrow -B} + N_{-A \rightarrow B})}{N * N} \quad (2.7)$$

$$U_4 = \frac{(N_{A \rightarrow B} + N_{-A \rightarrow B}) * (N_{-A \rightarrow B} + N_{-A \rightarrow -B})}{N * N} \quad (2.8)$$

$$del_1 = 1 - \frac{N_{A \rightarrow -B}}{U_1 * N} \quad (2.9)$$

$$del_2 = 1 - \frac{N_{A \rightarrow B}}{U_2 * N} \quad (2.10)$$

$$del_3 = 1 - \frac{N_{-A \rightarrow -B}}{U_3 * N} \quad (2.11)$$

$$del_4 = 1 - \frac{N_{-A \rightarrow B}}{U_4 * N} \quad (2.12)$$

The first four values of  $del_p$  relate to each of the unsymmetrical implications. The values of  $del_p$  for symmetrical implications can be found by combining values of two each of the unsymmetrical implications as shown below.

$$del_5 = \frac{U_1 * del_1 + U_4 * del_4}{U_1 + U_4} \quad (2.13)$$

$$del_6 = \frac{U_2 * del_2 + U_3 * del_3}{U_2 + U_3} \quad (2.14)$$

The implications are associated with two weight functions that specify the strength of the relationship between the pair of nodes that are connected. These weight functions can thus be defined as shown below [1, 2].

$$W_I : N_{pre} \times N_{con} \rightarrow [0,1] \quad (2.15)$$

$$\overline{W}_I : -N_{con} \times -N_{pre} \rightarrow [0,1] \quad (2.16)$$

Thus they can be defined in terms of the contents of the contingency table as below.

$$W_I = \frac{N_{A \rightarrow B}}{N_{A \rightarrow B} + N_{A \rightarrow -B}} \quad (2.17)$$

$$\overline{W}_I = \frac{N_{-A \rightarrow -B}}{N_{-A \rightarrow -B} + N_{A \rightarrow -B}} \quad (2.18)$$

Thus if  $\hat{I}$  is a complete set of possible implication rules which can be generated,  $R$  is the relation type,  $W_I$  and  $\overline{W}_I$  are the weight functions that map the precedent node  $N_{pre}$  and the consequent node  $N_{con}$ , then an implication rule can be generalized as follows [2].

$$I \in \hat{I}, I = \langle R, N_{pre}, N_{con}, W_I, \overline{W}_I \rangle \quad (2.19)$$

Implication networks extract many more relationships among the variables that are overlooked by most of the current approaches. Most of the currently existing approaches concentrate only on the relations which have same states for both the variables like in the cases of positive equivalence and negative equivalence. But there might be some very significant connections in the implication networks which are not significantly correlated. The implication networks have the capability to identify many known biological phenomena and also to extract hierarchical relationships. They are also stable over various species.

## 2.4 Survival Analysis

To validate the prognostic signatures identified in the network-based approach, survival analysis is performed on them. Survival Analysis is normally done with respect to the occurrence of an



event (which is normally death of the patient) with time. It helps in finding out what portion of the considered group survives past a certain time. It also gives the rate of increase or decrease of the occurring event. Survival Analysis is done using the following methods

1. Time dependant ROC analysis and Random Test
2. Cox Proportional Hazards Model
3. Kaplan-Meier Plots
4. Multivariate Analysis using Cox Proportional Hazards Model

#### **2.4.1 Time dependent ROC analysis and Random Test**

ROC curves are techniques used for visualization, organization, and selection of classifiers based on their performance. They are plots between sensitivity and (1-specificity). Time dependent ROC analysis is said to be done when ROC (Receiver Operating Characteristics) curves are varied as functions of time  $t$ . Since most of the disease outcomes are dependent on time, Time dependent ROC analysis becomes more apt.

ROC curves are capable of portraying the differentiation capacity of a test even without considering a specific threshold [16]. Even when the diagnostic markers are on diverse scales of measurement, ROC curves provide a convenient method for comparison. AUC or the Area under the Curve [17] is also considered as an important standard of comparison. It can be considered as the metric that compares the probability of diseased states to non-diseased states and thus summarizes the ROC curve. Since ROC is a two-dimensional representation of the classifier performance, it can be reduced to a single scalar value as AUC representing the expected performance. Thus realistic classifiers should not have AUC value less than 0.5. Since ROC graphs are conceptually very simple, they can be used as cost-sensitive learning techniques.

If  $X$  is the explanatory variable or predictor, and  $D(t)$  is the event (which is death in our case) at any time  $t$ . If a cutoff point  $c$  is considered which keeps varying, then the sensitivity and specificity would be functions of  $c$  and  $t$ . Thus the sensitivity and specificity can be expressed as [16]

$$\text{sensitivity}(c, t) = P\{X > c | D(t) = 1\} \quad (2.20)$$

$$\text{specificity}(c, t) = P\{X \leq c | D(t) = 0\} \quad (2.21)$$

Thus  $ROC(t)$  curve is a graph plotted between  $\text{sensitivity}(c, t)$  (it is the Y-axis in the ROC curve) and  $\{1 - \text{specificity}(c, t)\}$  (it is the X-axis in the ROC curve). The area under the curve for each  $ROC(t)$  is defined as  $AUC(t)$ .

In Random Test, gene signatures are picked randomly and their performance is compared with the performance of our gene signature. Thus it acts as a measure of the significance of our signature when compared to some signatures picked randomly.

#### **2.4.2 COX Proportional Hazards model**

Cox Model was a regression model described by D.R.Cox in his paper, “Regression Models and Life-Tables” [18] in 1972. Since then till to date, Cox model is a well-recognized statistical technique which explores the relationship between survival times and several other predictors (also called covariates or explanatory variables) simultaneously [18]. In other words, it gives an estimate of the treatment effect on the survival after adjusting the covariates and also to estimate the risk of death. Cox model has many coefficients. For each variable these coefficients describe whether a patient is under poor prognosis or a good prognosis.

Cox model helps in isolating the effects of treatment from the effects of other variables or covariates. It helps in improving the treatment effect as it narrows down the confidence interval. Survival times are censored if the patients followed up for several years are still alive after the end of study. Their survival time is not known from their surgery, as it is even longer than the time in study.

The regression model introduced by Cox is also known as proportional hazards regression analysis as it is used to explore several variables at a time. The hazard function is the probability that a patient will experience an event within a small interval of time, and therefore it can be understood as the risk of dying at time  $t$ . The hazard function denoted by  $h(t)$  can be estimated using the equation [19]. If two observations are considered as shown below, the hazard ratio of these two observations is shown in the last equation.

$$h(t) = \frac{\text{number of individuals experiencing an event in interval beginning at } t}{(\text{number of individuals surviving at time } t) \times (\text{interval width})} \quad (2.22)$$

$$h_i(t) = h_0(t) \times \exp(\beta_1 \cdot x_{i1}) \quad (2.23)$$

$$h_j(t) = h_0(t) \times \exp(\beta_1 \cdot x_{j1}) \quad (2.24)$$

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\beta_1 \cdot x_{i1})}{\exp(\beta_1 \cdot x_{j1})} \quad (2.25)$$

### 2.4.3 Multivariate Analysis using COX Proportional Hazards model

Regression is a statistical technique used to explain the relationship between the values of two or more variables. When more than one variable needs to be taken into account, the method is called multiple regression technique (multivariate analysis) which is almost the same as Cox's

model except that Cox model allows considering in to account more than one explanatory variable at any one time. Thus hazard [19] at any time  $t$  can be expressed as

$$h(t) = h_0(t) \times \exp (\beta_{\text{age}} \cdot \text{age} + \beta_{\text{duration}} \cdot \text{duration} + \dots + \beta_{\text{location}} \cdot \text{location}) \quad (2.26)$$

By applying natural logarithms, we get

$$\ln h(t) = \ln h_0(t) + \beta_{\text{age}} \cdot \text{age} + \beta_{\text{duration}} \cdot \text{duration} + \dots + \beta_{\text{location}} \cdot \text{location} \quad (2.27)$$

Thus  $h_0(t)$  is the underlying hazard function or baseline hazard. The coefficients such as  $\beta_{\text{age}}, \beta_{\text{duration}}, \dots, \beta_{\text{location}}$  are the regression coefficients and they constitute the proportional change that can be expected in the hazard or risk function related to the other variables which are estimated by a statistical method called the maximum likelihood technique.

Consider two observations of hazards at times  $i$  and  $j$ .

$$h_i(t) = h_0(t) \times \exp (\beta_{\text{age}} \cdot \text{age}_i + \beta_{\text{duration}} \cdot \text{duration}_i + \dots + \beta_{\text{location}} \cdot \text{location}_i) \quad (2.28)$$

$$h_j(t) = h_0(t) \times \exp (\beta_{\text{age}} \cdot \text{age}_j + \beta_{\text{duration}} \cdot \text{duration}_j + \dots + \beta_{\text{location}} \cdot \text{location}_j) \quad (2.29)$$

The hazard ratio for the above two would be [19]

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp (\beta_{\text{age}} \cdot \text{age}_i + \beta_{\text{duration}} \cdot \text{duration}_i + \dots + \beta_{\text{location}} \cdot \text{location}_i)}{\exp (\beta_{\text{age}} \cdot \text{age}_j + \beta_{\text{duration}} \cdot \text{duration}_j + \dots + \beta_{\text{location}} \cdot \text{location}_j)} \quad (2.30)$$

Thus it can be seen that the hazard ratio does not depend on the baseline hazard. Proportional hazard is the assumption of a constant relationship between the dependent variable and the explanatory variables. Cox regression analysis will result in a final model which yields an equation for hazard as a function of the several covariates.

#### 2.4.4 Kaplan-Meier Plots

The Kaplan-Meier curves, which are also known as product limit estimators, help in estimating the survival function from life time data.

From a set of survival times, the proportion of the population who would survive a given length of time under the same circumstances can be estimated using the Kaplan-Meier [48] (or product limit) method. A plot of the Kaplan-Meier estimate of the survival function is a step function. The estimated survival probabilities are constant between adjacent death times and they only decrease at each death.

If  $n_i$  is the number of samples at risk just prior to time  $t_i$  and  $d_i$  is the number of deaths at time  $t_i$ , then Kaplan-Meier estimate would be the nonparametric maximum likelihood estimate of  $S(t)$ , which is the probability that a sample from the given population would have a lifetime exceeding  $t$ , which can be shown as below.

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \quad (2.31)$$

## 2.5 Prognostic Validation

Prognostic validation is usually done to predict the chance of recovery of a patient. Prognosis is normally estimated with the help of variables such as sensitivity, specificity, hazard ratios and log-rank p-values. Prognostic validation is done using the following methods

1. Overall Accuracy
2. Concordance Probability Estimate
3. Gene Set Enrichment Analysis

### 2.5.1 Overall Accuracy

There are many metrics such as likelihood ratio, area under receiver operator characteristic (ROC) curve, overall accuracy, etc. which integrated both sensitivity and specificity to describe the validity of models or tests. Overall Accuracy is a single summary metric which is calculated from the  $2 \times 2$  contingency tables which is the overall probability that a patient will be correctly

		Disease		Sensitivity = $\frac{a}{a+c}$
		Positive	Negative	
Test	Positive	a	b	Prevalence = $\frac{a+c}{N}$
	Negative	c	d	
		$a+c$	$b+d$	
		$N = a + b + c + d$		

classified by a model. A  $2 \times 2$  contingency table is shown below.

$$\text{Accuracy} = \frac{a+d}{N} = \left(\frac{a+c}{N}\right) \left(\frac{a}{a+c}\right) + \left(\frac{b+d}{N}\right) \left(\frac{b}{b+d}\right) \quad (2.32)$$

$$\therefore \text{Accuracy} = (\text{Prevalence}) (\text{Sensitivity}) + (1-\text{Prevalence}) (\text{Specificity}) \quad (2.33)$$

Sensitivity is the probability that a person with the disease tests positive. Specificity is the probability that a disease-free person tests negative. Disease prevalence refers to the ratio of the number of patients with the disease and the total number of patients considered. Overall accuracy is the probability that a patient tests positive when he has the disease and tests negative when he is disease free; that is, the sum of true positives and the true negatives divided by the total number of patients. In other terms it can be shown to be the weighted average of the sensitivity and specificity where sensitivity is weighted by prevalence and specificity is weighted by the complement of prevalence. Thus the formulae for Sensitivity, Specificity and Overall Accuracy when a  $2 \times 2$  contingency table is considered are shown above.

Overall accuracy as a measure has a limitation in the sense that it is prevalence dependent which sometimes gives a misrepresented idea of the validity of the model.

### 2.5.2 Concordance Probability Estimate (CPE)

Concordance probability is used to estimate the distinguishing power and the predictive accuracy of statistical models. CPE forms a stable estimator of predictive accuracy which can be computed easily. The proposed estimator for CPE is a function of regression parameters and the covariate distribution and is asymptotically unbiased. A concordance probability of 1.0 represents a model that has perfect discriminating capacity where as a CPE of 0.5 indicates that the model is not good enough as it cannot discriminate between the observations in an accurate manner. If two observations  $(X_1, T_1)$  and  $(X_2, T_2)$  are considered, then their concordance probability [29] is defined as below.

$$CPE_{X,T} = pr(T_2 > T_1 | X_2 \geq X_1) \quad (2.34)$$

If the value of CPE is less than 0.5, it does not mean that the model is bad, but it may be considered as below by taking  $-X$  instead of  $X$  as a predictor of  $T$  to obtain a CPE higher than 0.5.

$$1 - CPE_{X,T} = pr(T_1 > T_2 | X_2 \geq X_1) = CPE_{-X,T} \quad (2.35)$$

If  $x$  is a p-dimensional covariate vector, and  $h_0(t)$  is the baseline hazard function independent of the covariates, and  $\beta_0^T$  is the vector of the regression parameters, then the hazard function  $h(t|x)$  of Cox proportional hazards model is given by

$$h(t|x) = h_0(t) \exp(\beta_0^T x) \quad (2.36)$$

CPE is a simple function of the Cox proportional hazards model and is not sensitive to the degree of censoring and does not require imputation of survival times.

The proportional hazards conditional survival function which determines the relationship between the  $p$ -dimensional covariate vector  $x$  and the survival time  $t$  is given by

$$S(t; x, \beta) = \exp \left\{ -\exp(\beta^T x) \int h_0(t) dt \right\} \quad (2.37)$$

The ordering between the survival times of two subjects with log relative risks  $\beta^T x_1$  and  $\beta^T x_2$  under proportional hazards can be measured by

$$pr\{T(\beta^T x_2) > T(\beta^T x_1)\} = \int_0^\infty S(t; x_2, \beta) dS(t; x_1, \beta) = \frac{1}{1 + \exp\{\beta^T(x_2 - x_1)\}} \quad (2.38)$$

Thus concordance probability  $CPE(\beta) = pr(T_2 > T_1 | \beta^T x_1 > \beta^T x_2)$  may be written as

$$\frac{\iint [1 + \exp\{\beta^T(x_2 - x_1)\}]^{-1} dF(\beta^T x_1) dF(\beta^T x_2)}{\iint dF(\beta^T x_1) dF(\beta^T x_2)} \quad (2.39)$$

for integrals ranging over the interval  $\beta^T x_1 > \beta^T x_2$  and  $F$  is the distribution function of the covariate linear combination  $\beta^T X$ .

The concordance probability estimator [29] can be given as

$$CPE_n(\hat{\beta}) = \frac{2}{n(n-1)} \sum_{i < j} \sum \left\{ \frac{I(\hat{\beta}^T x_{ji} < 0)}{1 + \exp(\hat{\beta}^T x_{ji})} + \frac{I(\hat{\beta}^T x_{ij} < 0)}{1 + \exp(\hat{\beta}^T x_{ij})} \right\} \quad (2.40)$$

where  $\hat{\beta}$  is the partial likelihood estimator for  $\beta$  and empirical distribution function was used for and  $x_{ij}$  represents the pairwise difference  $x_i - x_j$ .



In R, a package for CPE exists which includes a command named “phcpe”. This command is a function used to calculate the Gonen & Heller concordance probability estimate (CPE) for the Cox proportional hazards model. It outputs the CPE and the standard error of the CPE. The input for phcpe is a Cox fit model. Since a Cox fit model is present, we get various outputs such as p-values from log-rank tests, hazard ratios and the confidence intervals also as outputs from phcpe. The CPE values must always be greater than 0.5 for a data to be significant. The higher the CPE values, the more significant the data is considered to be. Similarly, p-values from log-rank tests must be lesser than 0.05 and hazard ratios must be greater than at least 1.

### **2.5.3 Gene Set Enrichment Analysis (GSEA)**

Gene Set Enrichment Analysis<sup>12</sup> is a powerful analytical method that computes whether a predefined set of genes is statistically significant and whether the gene set has agreeable differences between the two phenotypes (biological states). GSEA interprets gene expression data and focuses on the gene sets as a whole. It generates analysis based on the groups of genes that share common biological function, chromosome location or regulation.

GSEA [30] works in three steps:

1. Calculation of an Enrichment Score (ES): ES is the maximum deviation from zero encountered in a random walk which corresponds to the Kolmogorov-Smirnov-like statistic. It is calculated to show whether the gene set is overrepresented at the top or bottom of the ranked list.
2. Estimation of the significance level of ES: The nominal P value of the enrichment score which denotes its statistical significance is estimated by using an empirical phenotype-based permutation test. The permutation of phenotypes preserves the complex correlation structure

---

12. <http://www.broadinstitute.org/gsea/>

of the gene expression data which provides a more biologically reasonable estimation of significance.

3. Adjustment for Multiple Hypothesis Testing: After the complete database of gene sets is assessed, the estimated significance level is adjusted to account for multiple hypothesis testing by calculating the normalized enrichment score (NES) and the false discovery rate (FDR). FDR is the estimated probability that a gene set with a given NES represents a false positive finding.

## 2.6 Topological Validation

Structural validation is done using the following methods

1. PRODISTIN<sup>13</sup>
2. PubMed<sup>14</sup>
3. NCI Pathways<sup>15</sup>
4. KEGG<sup>16</sup>
5. MATISSE<sup>17</sup>
6. STRING 8<sup>18</sup>
7. Ingenuity Pathway Analysis<sup>19</sup>
8. Pathway Studio<sup>20</sup>

### 2.6.1 PRODISTIN

PRODISTIN method functionally classifies genes or proteins from all types of interaction networks according to the identity of their interacting partners. It can also be used to obtain information related to protein function and to relationships linking cellular processes [22].

Proteins can be compared functionally at the cellular level or the molecular level. In

---

13. <http://crfb.univ-mrs.fr/webdistin/>

14. <http://www.genome.jp/kegg/>

15. <http://pid.nci.nih.gov/>

16. <http://www.ncbi.nlm.nih.gov/pubmed/>

17. <http://acgt.cs.tau.ac.il/matisse/>

18. <http://string.embl.de/>

19. <http://www.ingenuity.com/>

20. <http://www.ariadnegenomics.com/products/pathway-studio/>

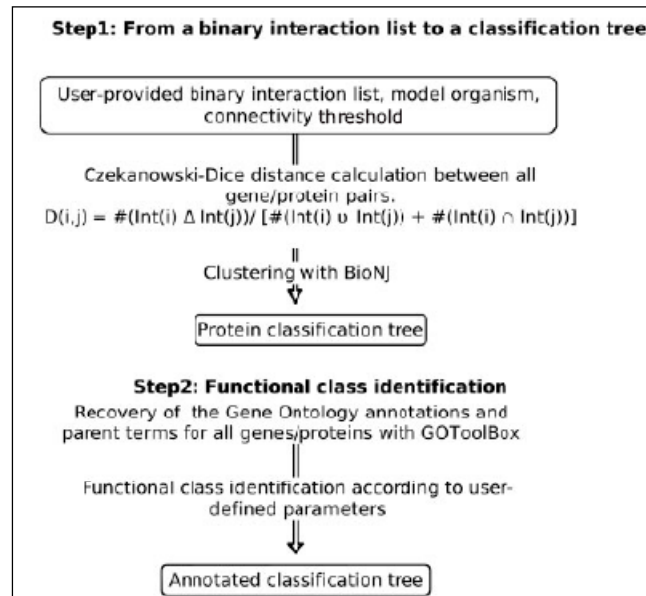
PRODISTIN method, proteins are clustered according to their cellular processes more efficiently rather than the molecular or the biochemical functions. The result of PRODISTIN method on interaction networks allows the user to acquire a classification tree in which network genes/proteins are grouped according to their functional similarity and to find out functional classes in the tree using the biological process Gene Ontology annotations of genes/proteins. The basic concept is that the more two proteins share interacting partners, the more they should be functionally related.

PRODISTIN method consists of two different and successive bioinformatic steps as shown in Figure 2-15 below. Initially a graph including all proteins connected by a specific relation is constructed and Czekanowski-Dice distance is calculated between all possible pairs of proteins in the graph with respect to the interacting partners they share. This classical distance on graphs corresponds to the formula [22].

$$D(i, j) = \frac{\#[Int(i) \Delta Int(j)]}{[\#[Int(i) \cup Int(j)] + \#[Int(i) \cap Int(j)]]} \quad (2.41)$$

where  $i$  and  $j$  denote two proteins,  $Int(i)$  and  $Int(j)$  are the lists of their interacting partners plus themselves (which are used to decrease the distance between the proteins interacting with each other) and  $\Delta$  is the symmetrical difference between the two sets. In other words it gives sum of the interactors in both minus twice the number of common interactors between the two interacting proteins. This distance was chosen because it increases the weight of the shared interactors by giving more importance to the similarities than to the differences and also it authorizes the use of tree representation. For two proteins that do not share any interactors, the distance is 1 and is the maximum value. For two proteins interacting with each other and sharing exactly the same interactors, the distance is 0 and is the minimum value. The second step would

be to cluster all the distance values according to BioNJ, which would lead to a classification tree. The tree can be visualized and subdivided in to formal classes according to the biological process in Gene Ontology annotations.



**Figure 2-15: Step wise procedure for PRODISTIN method [21]**

PRODISTIN has the ability to predict correctly, the function for unknown proteins and it shows reliability even in the presence of both spurious and missing interactions in the dataset. It can also be used to investigate the evolution of the function of duplicated genes. As more interactions become available, it improves the relevance of the protein clusters found by the PRODISTIN method. The PRODISTIN web interface is shown in Figure 2-16.

Prodistin Web Site: Functional classification of proteins based on interactions.		
<a href="#">Return</a> <a href="#">Home</a> <a href="#">Start</a> <a href="#">Help</a> <a href="#">References</a> <a href="#">Contact</a>		
<b>Step 1 : From a binary interaction list to a classification tree</b>		
Species	Drosophila melanogaster	Choose your favorite organism <a href="#">help</a>
List of interactions (demo)	<input type="text"/> <input type="button" value="Browse..."/>	Choose a file to upload <a href="#">help</a>
Gene/Protein connectivity	3	Minimal gene/protein connectivity to be classified <a href="#">help</a>
<input type="button" value="Compute"/>		

**Figure 2-16: PRODISTIN website**

## 2.6.2 PubMed

PubMed was developed by National Centre for Biotechnology Information (NCBI) at National Library of Medicine (NLM) located at U.S. National Institutes of Health (NIH). It is available through Entrez retrieval system. It helps search in biomedical citations and abstracts. PubMed is a search engine that allows access to many databases including the MEDLINE database of journal articles. Its focus is on medicine and related fields like nursing.

PubMed can be searched for required details in many ways as shown below in Figure 2-17. Any data required can be searched in entire PubMed or it can be restricted to some fields such as genes, proteins, journals, etc. The results could be even more focused by choosing the limits of search such as the organisms, taxonomy, etc. It is one of the web based search engines which is used widely by biostatisticians to extract gene information.

Advanced search is also available which allows finding data by author's name, publication date, title, etc. as shown below in Figure 2-18. The recent items searched are also stored and they can be revisited when required. PubMed has an option called LinkOut which allows to access resources in external websites directly from the PubMed database. LinkOut resources include

research tools, full text publications, biological databases, etc. Thus instead of searching various databases each separately, it would be sufficient to check in PubMed which give links to all the other databases if they are present. The links are present in such a way that we could access just the abstract or the entire text in required format (text or pdf).

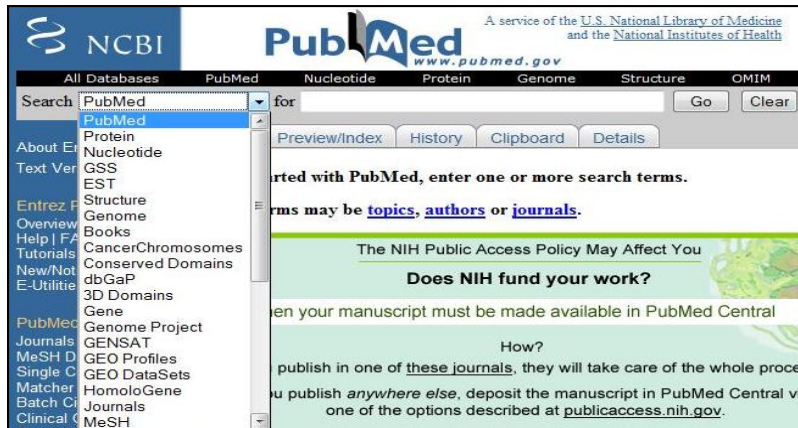


Figure 2-17: PubMed website

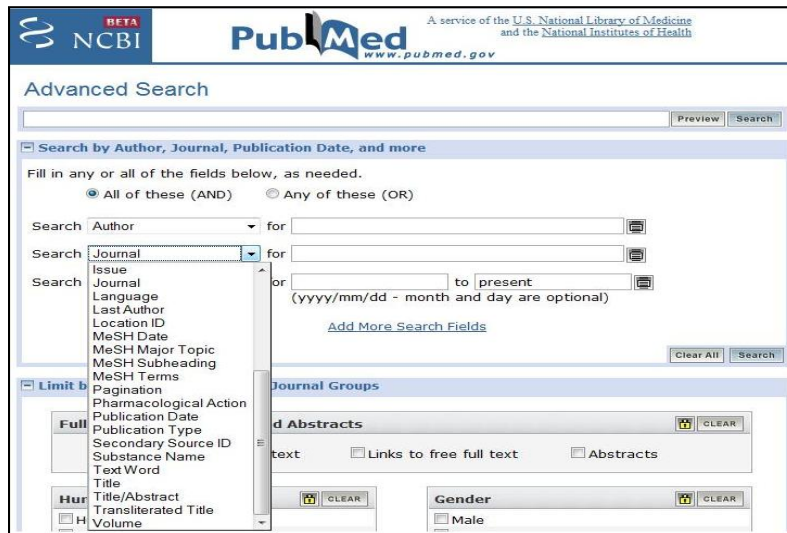



Figure 2-18: Advanced Search in PubMed

### 2.6.3 NCI Pathways (Pathway Interaction Database)

National Cancer Institute (NCI) is a part of the U.S. National Institutes of Health and supports a national network of cancer centers and supports research projects in cancer control. Nature Publishing Group (NPG) is a publisher of 60 prestigious scientific journals including the highly impact **Nature**, the international weekly journal of science. A collaborative project between NCI and NCI is the Pathway Interaction Database (PID).

PID is a highly structured database which includes a curated collection of information. The schema of the database is very flexible which makes it easy to store a wide range of information about cell signaling pathways. It includes known biomolecular interactions that are taking place in human cells and also includes key cellular processes which when combined make up signaling pathways. PID shows not only the predefined pathways but also interaction networks that are dynamically constructed. Since the editorial section of PID also includes outlines of recent research articles connected to cancer, it acts as a practical advice and tool to bioinformaticians and biologists.

[Jump to main content](#)   [Jump to navigation](#)



**nature** PathwayInteractionDatabase

---

[Home](#)

**Pathway Interaction Database**

- [Browse pathways](#)
- [Search database](#)
- [Advanced search](#)
- [Connected molecules](#)
- [Batch query](#)
- [Download data](#)
- [Curation calendar](#)

**Pathway updates**

- [Recently added pathways](#)
- [Research highlights](#)
- [Bioinformatics primers](#)

**Information**

**About us**

- [The Pathway](#)

**Pathway Interaction Database**

**Biomolecular interactions and cellular processes assembled into authoritative human signaling pathways**

- 
**79 Human Pathways** 4474 Interactions  
[Curated by NCI-Nature](#)
- 
**317 Human Pathways** 6469 Interactions  
 Imported from [BioCarta/Reactome](#)


**Search database**

Molecule name/ID or biological process term/ID

[Advanced search](#)  
[Connected molecules search](#)  
[Batch query](#)

**Figure 2-19: Pathway Interaction Database**

[Jump to main content](#)   [Jump to navigation](#)



**nature** PathwayInteractionDatabase

---

[Home](#) > [Browse pathways](#)

**Browse pathways**

If you don't find your pathway of interest either in the Browse Pathways list or the [Curation Calendar](#), please [Suggest a Pathway](#).

**Jump straight to the search source**

- [NCI-Nature curated pathways](#)
- [BioCarta imported pathways](#)
- [Reactome imported pathways](#)

NCI-NATURE CURATED Pathway	Source
<a href="#">ADP-ribosylation factor 1 pathway</a>	NCI-Nature curated
<a href="#">Alpha-synuclein signaling</a>	NCI-Nature curated
<a href="#">Alternative NF-kappaB pathway</a>	NCI-Nature curated
<a href="#">Angiopoietin receptor Tie2-mediated signaling</a>	NCI-Nature curated
<a href="#">Atypical NF-kappaB pathway</a>	NCI-Nature curated
<a href="#">BARD1 signaling events</a>	NCI-Nature curated
<a href="#">BCR signaling pathway</a>	NCI-Nature curated
<a href="#">BMD receptor signaling</a>	NCI-Nature

**Figure 2-20: Browsing Pathways in PID**



PID includes many pathways extracted from three types of data as shown in Figure 2-19 and Figure 2-20. They are NCI-Nature Curated data, BioCarta data and Reactome data. NCI-Nature Curated data are produced by Nature Publishing Group according to a few principles. They include Human Model System, Biological relevance, Authority and Consistent nomenclature. BioCarta was imported in to PID in June 2004 and Reactome was imported in December 2007. In NCI-Nature Curated data and Reactome data, biomolecules are annotated with Uniprot protein identifiers and relevant post-translational modifications whereas in BioCarta data, biomolecules are annotated by Entrez gene identifiers without post-translational modifications.

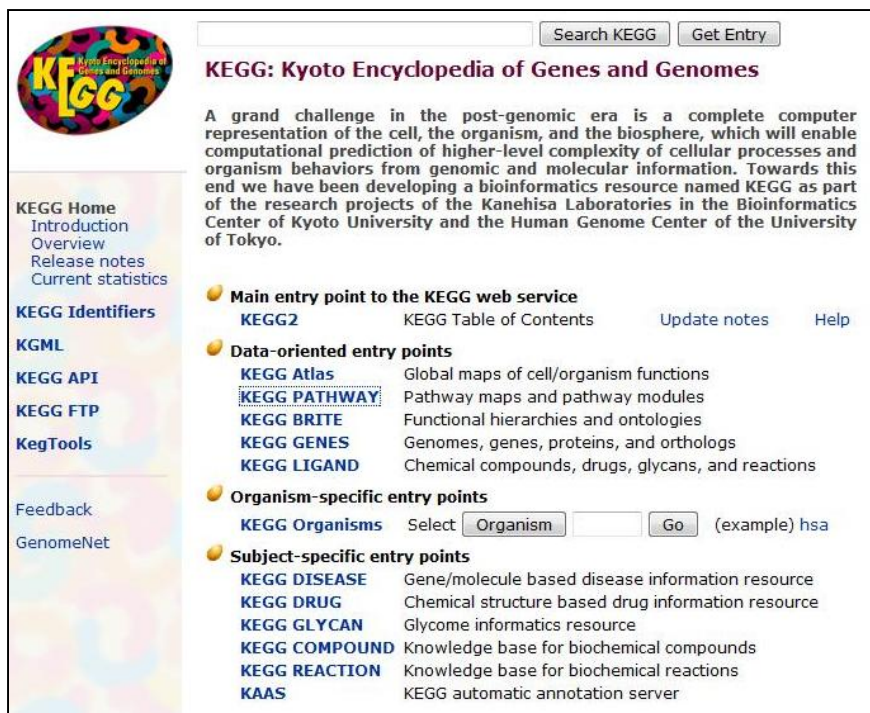
A biologically meaningful set of interactions is defined as a pathway in PID. Molecular interaction is the basic unit of representation in PID. Thus the information is very fine-grained and highly structured. In each interaction, each biomolecule is identified along with the nature of process (biological events) it is involved and its role in each of the processes. Pathways are portrayed graphically labeled nodes and edges. Additional references are also provided.

#### **2.6.4 KEGG**

KEGG stands for Kyoto Encyclopedia of Genes and Genomes. It was initiated in May 1995 to computerize the knowledge of molecular and cellular biology in terms of information pathways that consist of interacting genes or molecules. Its objective was to link [24] the structural data obtained by genome projects and the functional data. KEGG was built based on the pair wise interaction of genes or molecules. Since information regarding known pathways has been expanding rapidly, it has become necessary to computerize known pathways at the time of KEGG's initiation. KEGG was considered to be an effort to advance concepts and technologies and real time data collection efforts [26]. KEGG contains an aspect of the deductive database where new interactions could be deduced from relations stored in database. Thus the basic

concept of KEGG is the relation and deduction. KEGG has a hierarchy which is important in the sense that it represents functional, structural and evolutionary relationships of genes and molecules. The advancements in the database and networking technology make KEGG even better in the aspect of its functionality especially the deductive and object-oriented databases.

KEGG consisted of three databases [23, 25] when it started. KEGG PATHWAY represents higher order functions in terms of the network of interacting molecules (mostly proteins). It is a set of manually drawn pathway maps which represent knowledge on the molecular interaction and reaction networks and also on structural relationships. The best organized part of PATHWAY is that the organism specific pathways are constructed computationally by correlating genes in the genome with gene products in the reference pathways according to the matching EC numbers. Gene catalogs for all the completely sequenced genomes and some partial genomes are accumulated under KEGG GENES. The number of GENES's entries keeps increasing every year to keep track of the updating genome sequences. Thus GENES acts as a gateway to a number of other resources containing more detailed information. KEGG LIGAND is the collection of chemical compounds in the cell, enzyme molecules and enzymatic reactions. It is a composite database which includes COMPOUND, DRUG, GLYCAN, REACTION, RPAIR and ENZYME databases. KEGG BRITE was added later which depicts the hierarchical classifications.



**KEGG: Kyoto Encyclopedia of Genes and Genomes**

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

**Main entry point to the KEGG web service**  
[KEGG2](#)    [KEGG Table of Contents](#)    [Update notes](#)    [Help](#)

**Data-oriented entry points**  
[KEGG Atlas](#)    Global maps of cell/organism functions  
[KEGG PATHWAY](#)    Pathway maps and pathway modules  
[KEGG BRITE](#)    Functional hierarchies and ontologies  
[KEGG GENES](#)    Genomes, genes, proteins, and orthologs  
[KEGG LIGAND](#)    Chemical compounds, drugs, glycans, and reactions

**Organism-specific entry points**  
[KEGG Organisms](#)    Select   (example) hsa

**Subject-specific entry points**  
[KEGG DISEASE](#)    Gene/molecule based disease information resource  
[KEGG DRUG](#)    Chemical structure based drug information resource  
[KEGG GLYCAN](#)    Glycome informatics resource  
[KEGG COMPOUND](#)    Knowledge base for biochemical compounds  
[KEGG REACTION](#)    Knowledge base for biochemical reactions  
[KAAS](#)    KEGG automatic annotation server

Figure 2-21: KEGG Website



**KEGG PATHWAY Database**  
 Wiring diagrams of molecular interactions, reactions, and relations

[KEGG2](#)    [ATLAS](#)    [PATHWAY](#)    [BRITE](#)    [GENES](#)    [SSDB](#)    [LIGAND](#)    [DBGET](#)

**Pathway Maps**  
**KEGG PATHWAY** is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

- 1. Metabolism**  
 Carbohydrate    Energy    Lipid    Nucleotide    Amino acid    Other amino acid  
 Glycan    PK/NRP    Cofactor/vitamin    Secondary metabolite    Xenobiotics
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

- 6. Drug Development**

**Pathway Modules**  
**KEGG MODULE** is a new collection of pathway modules, molecular complexes, and other functional units, each represented as a list of KEGG Orthology (KO) identifiers. KEGG MODULE can be accessed through a BRITE hierarchy:  
**KEGG pathway modules**  
 or by the DBGET search.

Search  for

b

Figure 2-22: KEGG PATHWAY

KEGG was later updated with many other databases which came up to 19 databases categorized in to systems information, genomic information and chemical information. KEGG website can be seen in Figure 2-21 and KEGG PATHWAY is shown in Figure 2-22 above.

## 2.6.5 MATISSE

MATISSE is a program that implements a novel computational method for efficient detection and analysis of JACs. JACs are Jointly Active Connected Subnetworks which are the functional modules that are sought by identifying the connected subnetworks in the interaction data that exhibit high average internal similarity [27]. MATISSE has a statistical basis, which allows confidence estimation of the results and no prior knowledge of the JACs is required which removes the requirement of precalculation of the statistical significance of expression values. Thus it suits all types of network data overlaid with pair wise similarities.

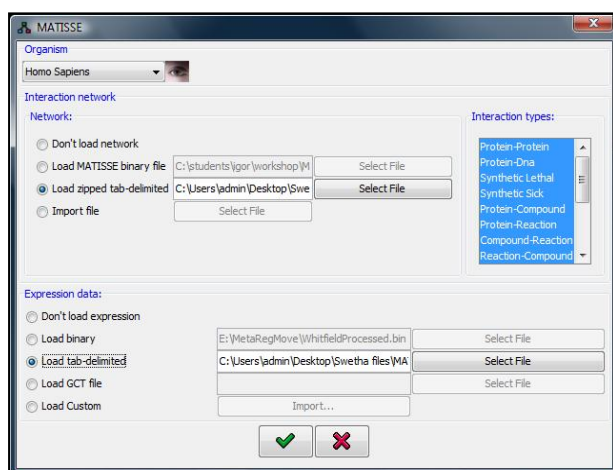


Figure 2-23: MATISSE interface

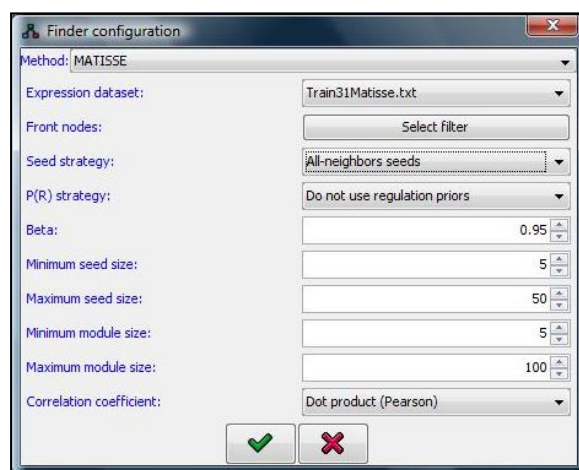
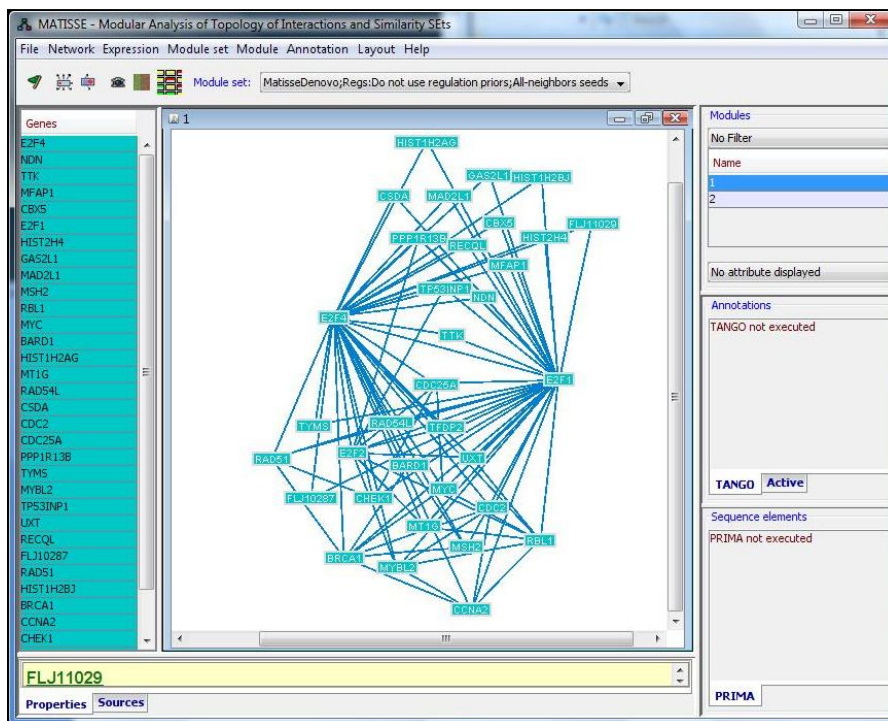


Figure 2-24: Choosing the Algorithm



**Figure 2-25: Displaying the module**

MATISSE needs to choose a species first and then the interaction network and the expression data must be loaded as shown in Figure 2-23. After the data is loaded, modules must be found and for that the Algorithm which must be used to find the modules must be selected as shown in Figure 2-24. The displayed modules can be filtered by applying filters which is shown in the Figure 2-25.

MATISSE detects non-overlapping JACSs by identifying heavy subgraphs in an edge-weighted similarity graph while maintaining connectivity in the interaction network. There are three phases in the detection and analysis of JACSs: (1) relatively small, high-scoring gene sets, called seeds must be detected; this detection can be done by any of the methods such as Best-neighbors or All-neighbors or Heaviest-subnet, (2) Seed improvement or greedy optimization; this optimization can be done using the methods like Node addition, Node removal, Assignment

change, and JACS merge, (3) significance-based filtering; the empirical p-value of the score was calculated for each JACS and a threshold of  $p=0.05$  is applied.

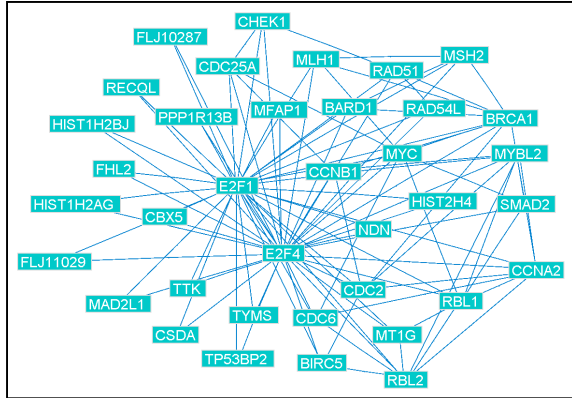


Figure 2-26: Module from MATISSE

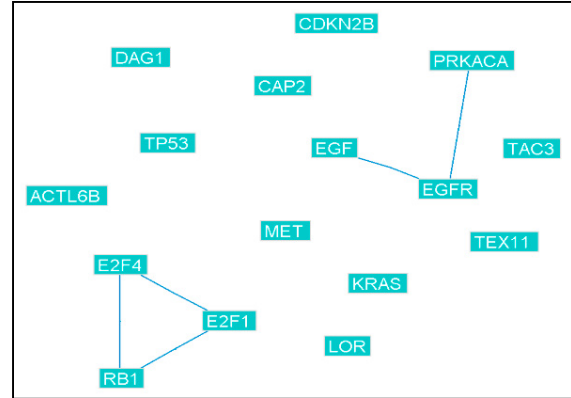


Figure 2-27: Module from Co-clustering method

In MATISSE, modules are also generated using Co-clustering method for the purpose of comparison. Co-clustering methodology uses a distance function that combines similarity of gene expression profiles with network topology. A few properties include Expression homogeneity (calculated as the Pearson correlation between genes within the same module), edge density (number of edges it contains as a fraction of all its node pairs), and clustering coefficient (fraction of a node's neighbor pairs connected in the network). MATISSE is designed to produce connected subnetworks as shown in Figure 2-26 whereas Co-clustering generates modules that are highly disconnected as shown in Figure 2-27 above. Thus MATISSE is much better [27] in all the properties checked for comparison with Co-clustering technique.

### 2.6.6 STRING 8

STRING 8 stands for eighth version of Search Tool for the Retrieval of Interacting Genes/Proteins. It is a database and web resource that constitutes most of the available protein-protein interactions, scores and weighs it and escalates it with not only predicted interactions, but

also with the results of automatic literature-mining searches. The latest version of STRING 8 covers almost 2.5 million proteins from 630 organisms and thus it provides a very comprehensive view of the protein-protein interactions. It includes resources from various other sources such as MINT, HPRD, BIND, DIP, BioGRID, KEGG, Reactome, IntAct, EcoCyc, NCI-Nature Pathway Interaction Database, and Gene Ontology (GO) protein complexes, etc. Apart from the interactions previously known interactions from the above resources, STRING 8 uses a number of prediction algorithms that computationally predict many more interactions. It searches for genes that are found in close surroundings of chromosomes as it would be a good indication of functional relation. It then searches for instances where genes join to encode a single fusion protein. It also searches for gene families that have similar phylogentic profiles and also genes that are co-expressed under different conditions. It includes interactions identified from text mining of databases like SGD, OMIM, The Interactive Fly and all the abstracts of PubMed. STRING 8 is more fault tolerant when clustering conserved neighborhoods by ignoring false predictions.

The network images in STRING are generated using a spring model where nodes are taken as masses and the edges are considered as springs.

In STRING 8, one or more proteins of interest are entered as inputs by giving names or identifiers as shown in Figure 2-28. The appropriate organism is selected. STRING 8 also has a random input generator which will select randomly a gene/protein with a minimum of 4 predicted links above medium confidence or even better. Prediction summary is obtained for the proteins that were given as input. All the predicted relations are sorted by their scores and each

of them can be viewed in detail. The various types of evidence supporting the predicted associations can be viewed by clicking the different views of the data.

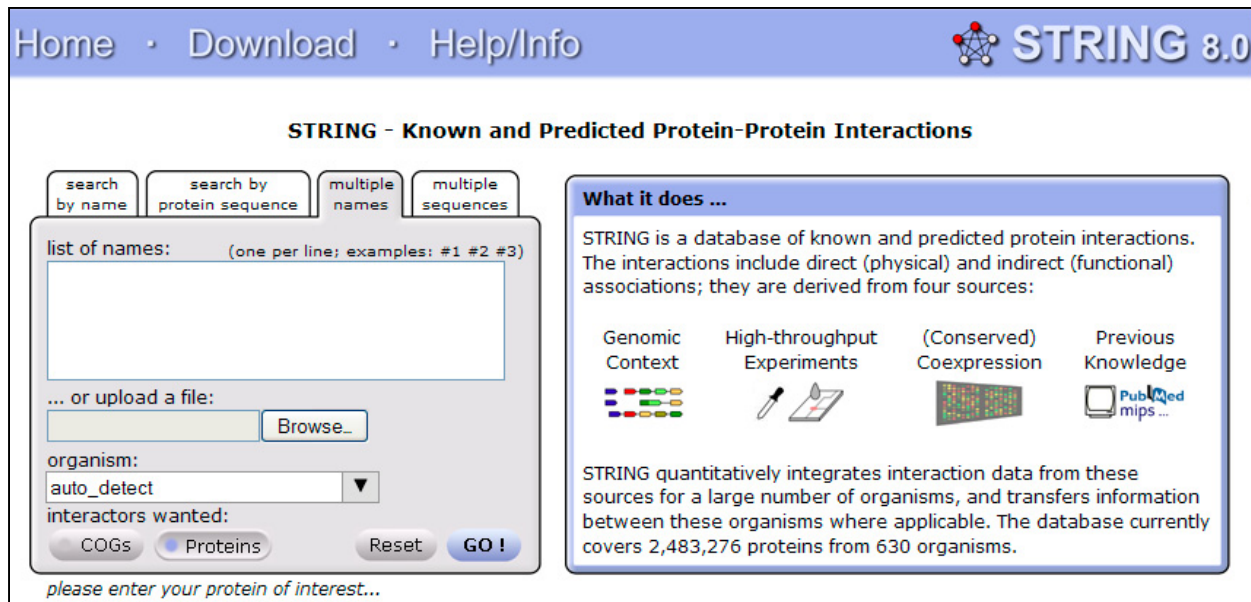


Figure 2-28: STRING 8 web interface

The network view briefly summarizes the interactions between the proteins with each of the protein as a node in the network. Any two proteins may be connected using seven different colored functional associations where each color indicates the presence of one evidence. A red line implies presence of fusion (individual gene fusion events per species) evidence; a green line implies conserved neighborhood (genes that occur repeatedly in close neighborhood in genomes) evidence; a blue line implies co-occurrence (presence or absence of linked proteins across various species) evidence; a purple line implies experimental (list of significant protein interaction datasets acquired from other protein-protein databases) evidence; a yellow line implies text mining (list of significant protein interaction groups extracted from the literature) evidence; a light blue line implies database (list of significant protein interaction groups acquired



from curated databases) evidence; a black line implies co-expression (genes that are co-expressed in the same species or other species) evidence. Clicking on a node gives many details about the protein and clicking the edge gives all the different scores relating to each of the evidences.

### **2.6.7 Ingenuity Pathway Analysis**

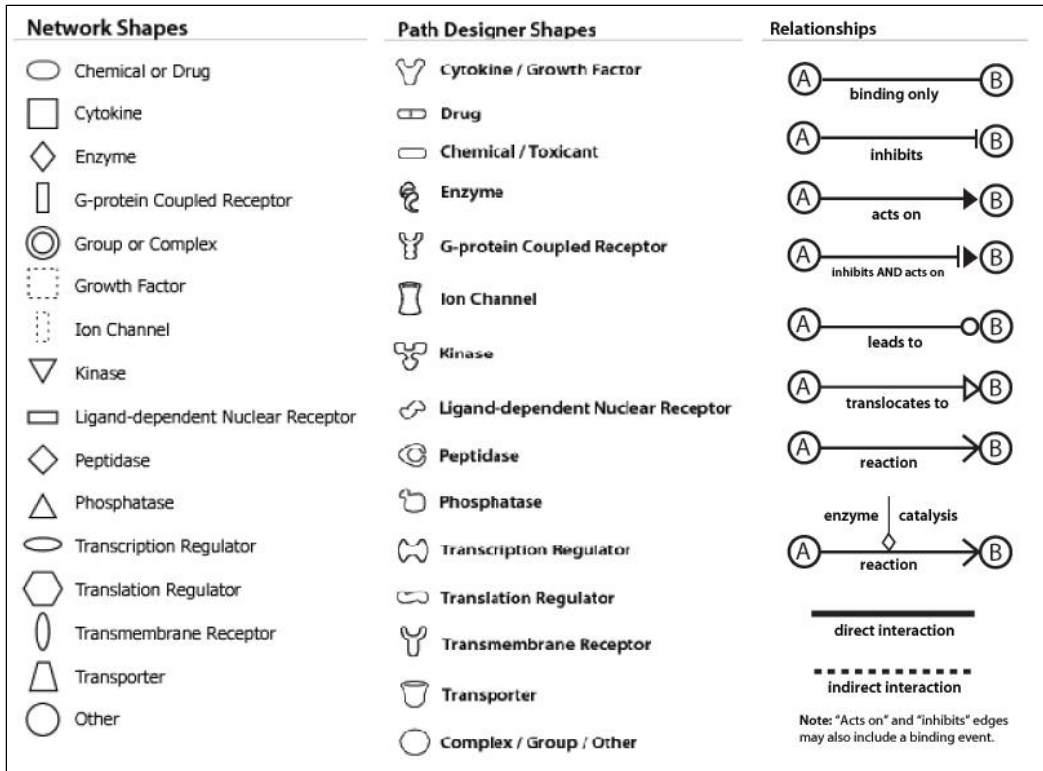
Ingenuity Pathway Analysis (IPA) integrates data from a variety of experimental platforms. It provides insight into molecular and chemical interactions, cellular phenotypes, and disease processes of our system. IPA is built upon a huge foundation of scientific evidence which include journal articles, textbooks, and other data sources. It presents the data in a meaningful visual and knowledgeable way.

There are different types of analyses that can be performed on a group of genes. They include Core analysis, IPA-Tox analysis, IPA-Metabolomics or IPA-Biomarker analysis. These analyses in most cases give a good indication of what cellular processes the given dataset is related to. Core analysis is used to interpret datasets in the context of biological processes, pathways and molecular networks. IPA-Metabolomics analysis analyzes the metabolite data about cell physiology and metabolism. IPA-Tox analysis assesses the toxicity and safety of the compounds of interest. It also shows the appropriate toxicity phenotypes and clinical pathology endpoints related to a dataset. IPA-Biomarker analysis identifies and prioritizes the most appropriate and promising molecular biomarker candidates from the datasets. Each of these analyses can be run multiple times on different inputs and they can be compared among themselves using Comparison analyses which help in understanding which of the samples are more relevant to each condition.

Ingenuity systems have a database that is highly structured and context-rich which makes it unique among the different pathway applications. All the results found in IPA are always supported by experimental results and thus are not just based on the occurrence in few abstracts. These results are structured in to an ontology which lets the use of very powerful computational algorithms that presents the results in IPA when queried.

IPA includes many features such as integrated broadband coverage of systems biology (including protein, gene, protein complex, cell, cellular component, tissue, organ, small molecule, and disease interrelationships), broad genome wide coverage of human, mouse, and rat genes, huge number of pathway interactions extracted from literature, very systematic capture of canonical pathway relationships and almost up to date literature.

IPA uses different shapes for the nodes in the networks and different types of connectors between the nodes for different types of relationships between the genes or proteins as shown in Figure 2-29.



**Figure 2-29: Different Network and Path Designer shapes along with Relationships used in IPA to represent different types of data**

It is not an open source and hence a license is needed to use the entire version of IPA. There is a free trial version but it does not include all the pathways and hence same results are not obtained each time it is used.

### 2.6.8 Pathway Studio

Pathway Studio is a combination of three products:

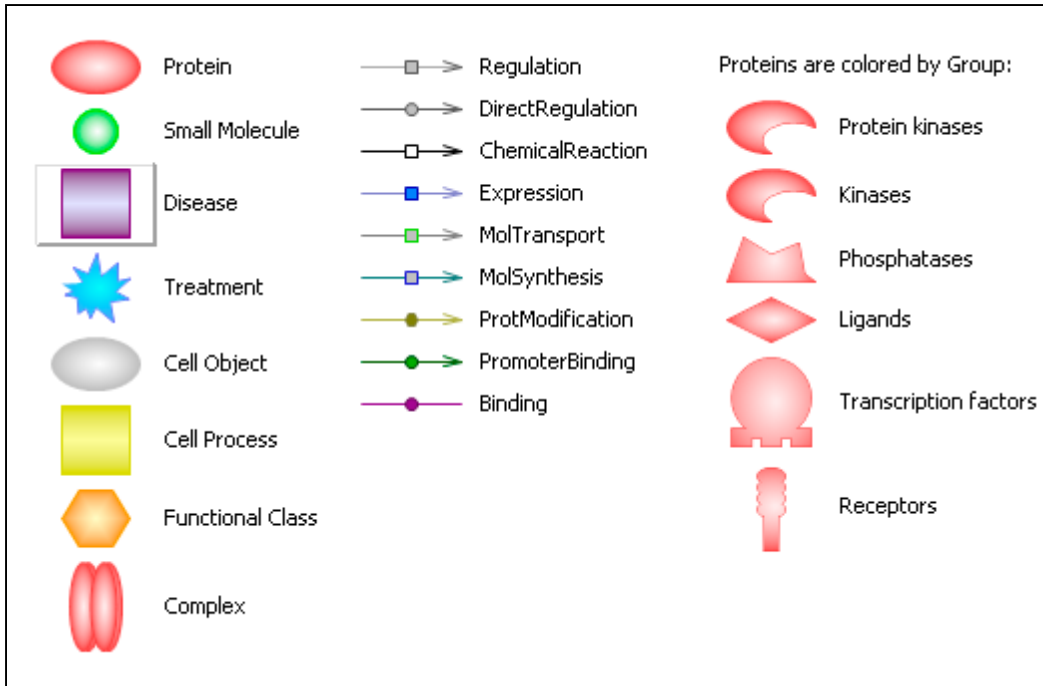
1. ResNet database
2. MedScan application
3. Pathway Studio interface

ResNet is a database that comprises of the biological relations, ontologies and pathways that were compiled by the Ariadne scientists for Mammalian (Human, Rat, and Mouse) and Plant

research. This database stores information that has been successfully extracted from PubMed in a manner such it allows searching, retrieving and even updating of the database by the user. Moreover the extracted interactions have access linked to the original data source. All possible aliases are also included which excludes redundancy and thus maintains identity of the genes.

MedScan is a computational approach used for data analysis that has the information from literature as a coherent and integrated part within it. It is like a web search engine which not only gathers knowledge about a query but also scans the literature for relationships and highlights those relationships in the articles that were gathered. It also lists all the relationships and molecular processes in appropriate tables which can be saved in to the ResNet database and reused for further analysis. MedScan has access to PubMed and 47 full-text journals and additional journals may also be added from different sources. MedScan can thus create many databases of specific organisms, diseases, etc and it can highlight different proteins, chemicals, cell processes, etc in literature. MedScan is used to update ResNet database and this can also be automated.

Pathway Studio is a software which analyses gene expressions and builds pathways. These pathways can be expanded and various relationships between genes, proteins, diseases, etc can be extracted. It works together with ResNet database (and MedScan reader to update the database). Once the experimental data is imported in to Pathway Studio, it enables in-depth analysis of the data and relationships are extracted from the literature (PubMed). By changing the settings in Pathway Studio, it can find common regulators and relates pathway components with biological entities of similar functionality.



**Figure 2-30: Different shapes and colors used in Pathway Studio to represent for different types of data**

There are 227 receptor signaling and 21 new cellular process regulation pathways that are included in Pathway Studio. These can be further expanded using MedScan. Pathway Studio is not an open source and license is needed to use it.

## 2.7 Summary

This chapter discussed the advantages of network-based techniques over feature selection methods. Feature selection procedures considered the behavior of each gene individually. Hence this chapter highlighted the importance of considering the “gene interactions” instead of each gene individually. It also reviewed the various network-based techniques such as correlation-based and clustering-based coexpression networks, Bayesian networks and artificial neural networks. The literature study of these techniques was summarized. The limitations of the

network-based techniques have been discussed. The correlation-based coexpression networks and the clustering based coexpression networks do not provide predictive structure or causal relations. Moreover their accuracy decreases with increase in network size. Bayesian networks cannot contain loops and they require subjective priors. Neural networks are time consuming and have the possibility of overtraining. It was thus discussed that all these limitations have been overcome by the implication networks. Algorithm of the implication network that was used in identifying the gene signatures was explained. Different validation techniques and tools, used for both prognostic and topological validation of the signature identified from the implication network were discussed.

### 3 Network-based approach for identification of prognostic signatures

#### 3.1. Introduction

In the earlier chapter, major limitations of the compared network techniques along with the advantages of implication networks were emphasized. In this chapter, details will be focused on how implication networks were applied on the dataset and how gene signatures were identified. Moreover descriptions of the datasets, procedures and results obtained from the application of implication networks are provided.

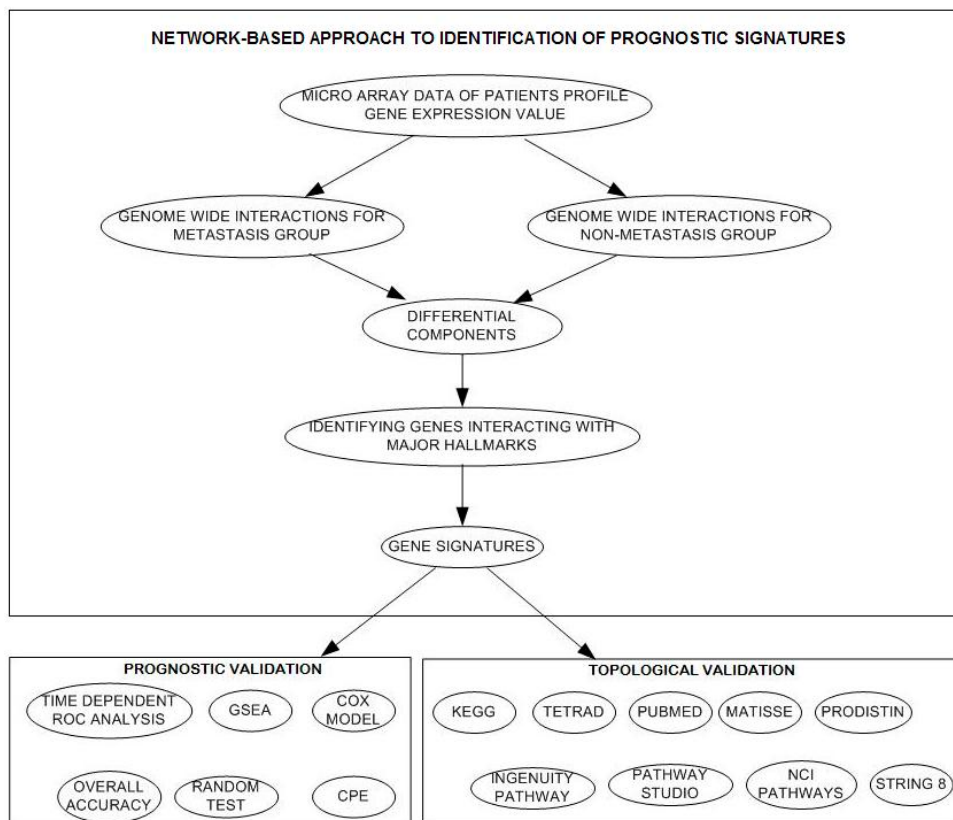


Figure 3-1: Flow chart of the methodology

The flow chart of the methodology is shown in the Figure 3-1 above. The first top box of the flow chart as displayed constitutes the identification of the gene signatures and this part will be discussed in this chapter. Details of survival analysis will also be given in this chapter.

### **3.2. Methodology for identifying prognostic gene signatures**

Identification of the gene signature is done on the training dataset and the validation of the obtained signature is performed on the test datasets.

The gene expression data of the training dataset is divided in to two or more groups based on some variables such as survival time, survival status, smoking status, etc. In this thesis, survival time and survival status are used together to split the data in to two groups. These groups are named Metastasis group (high risk group) and Non-Metastasis group (low risk group). Interactions among the genes are induced using prediction logic algorithm [1, 2]. Thus genome wide networks for both the groups are generated. After we get interactions among the genes in both the groups (Metastasis and Non-Metastasis), differential components between the groups are picked. Differential components are the set of interactions that are present in one group but are not present in the other group and vice versa. In other words, differential components of Metastasis group are unique and similarly, differential components of Non-Metastasis group are unique. Thus we get the interactions that differentiate the high risk group and the low risk groups.

Once the differential components are found, all the genes interacting with the major cancer hallmarks are picked. Major cancer hallmarks are genes that were already known to be of great importance in cancer research. Since these genes are known to have strong interactions with other genes in the cancerous conditions, we consider that the genes, which interact with all these



genes, might be in one or the other manner, included in the regulation or progression of tumors. Hence the set of genes that interact with all the major cancer hallmarks are considered to form a signature. By varying the set of hallmarks, we get different signatures.

Once signatures are found, we validate these signatures to find a signature that outperforms the other signatures and later evaluate the better signature prognostically and topologically (this will be dealt in the next chapter).

### **3.2.1 Datasets Information**

The data required for acquiring, training and testing the signatures was taken from a consortium [20] which was formed with the support and collaboration of US National Cancer Institute investigators. The dataset is a combination of samples collected from four institutions [20] using a common platform. The institutions that formed the consortium include University of Michigan Cancer Center (UM), Moffitt Cancer Center (HLM), Memorial Sloan-Kettering Cancer Center (MSK) and the Dana-Farber Cancer Institute (CAN/DF). The data from UM had 177 samples, HLM had 79 samples, CAN/DF had 82 samples and MSK had 104 samples. There were a total of 442 samples and 22215 genes. The data from UM and HLM were combined to form the Training dataset which has 256 samples. This Training dataset was used to find signatures and then they were validated on the remaining two datasets (CAN/DF and MSK).

### **3.2.2 Dataset processing**

The Training dataset consists of 22215 genes and 256 samples. There were duplicate probes for many genes. The duplicate probes of every single gene were averaged. This narrowed down the number to be 13658 unique genes. This data was split in to 2 files, Metastasis (High risk) and Non-Metastasis (Low risk) groups, based on the number of months the patients survived and

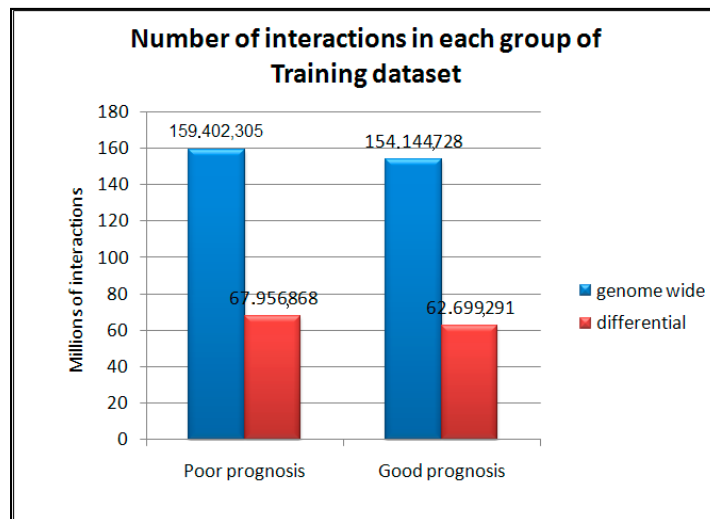
their survival status. If the number of months the patients survived was greater than 60 (5 years), the sample was put in Non-Metastasis group (Low risk). If the number of months the patients survived was less than 60 months and if it was known that the patient died, the sample was put in Metastasis group (High risk). If the number of months the patients survived was less than 60 months and if it was not known whether the patient died, the sample was censored. The Metastasis group (High risk) consisted of 125 patients and the Non-Metastasis group (Low risk) consisted of 104 patients. There were 27 samples that did not fall in to either of the groups and hence they were censored. Then the data values were converted in to binary values (0's and 1's) based on the mean of the expression values of each gene. The mean values were calculated using bootstrapping so that the impact of the number of samples being different does not affect the value of the mean. If the expression value of each gene for each sample was less than or equal to the mean of all samples for that particular gene, it was defined as 0 and if it was greater it was defined as 1.

### **3.2.3 Deriving genome scale gene interactions**

The interactions between the 13658 genes in each of the groups containing 125 (for metastasis group) and 104 (for non-metastasis group) samples were derived. The underlying principle in generating the gene interactions is based on prediction logic used for inducing the implication network [1, 2]. The minimum scope and minimum precision required were calculated using simple Z-test for a cutoff value of 1.64. There were 159,402,305 interactions that were derived from the Metastasis group (High risk). There were 154,144,728 interactions that were derived from the Non-Metastasis group (Low risk). The comparison of the number of interactions from both the groups is shown below in Figure 3-2.

### 3.2.4 Identifying Differential Components

After the genome wide interactions for the dataset were obtained, the differential components of each of the groups were obtained. Differential components are the gene interactions of one group that are not present in the other group. The interactions present in Metastasis group (High risk) but not present in the Non-Metastasis group (Low risk) are called the differential components of Metastasis group (High risk). Similarly, the interactions present in the Non-Metastasis group (Low risk) but not present in the Metastasis group (High risk) constitute the differential components of the Non-Metastasis group (Low risk). In other words they are the interactions that differentiate the two groups from one another. There were 91,445,437 common interactions between the groups. Thus there were 67,956,868 differential components for Metastasis group (High risk) and there were 62,699,291 differential components for Non-Metastasis group (Low risk) were. The comparison of the differential components from both the groups along with the genome wide interactions is shown below in Figure 3-2.



**Figure 3-2: Bar graph showing the number of interactions in Poor (high risk) and Good (low risk) prognosis for genome wide interactions and differential components in the Training dataset**

### **3.2.5 Major Cancer Hallmarks used to identify prognostic markers**

Major Cancer Hallmarks are the genes which are considered to be important. A different set of hallmarks can be considered to find different signatures. Gene signatures are picked from the genes which have interactions with all the hallmarks. There were 7 hallmarks that were considered to find gene signatures. Six of them were EGF, EGFR, KRAS, MET, RB1, and TP53. The seventh hallmark E2F had five different probes E2F1, E2F2, E2F3, E2F4, and E2F5. Hence they totaled to 11 hallmarks.

### **3.2.6 Identifying Gene Signatures**

All the genes which had interactions with the entire set of 11 hallmarks were picked from both the Metastasis group (High risk) and the Non-Metastasis group (Low risk). There were 7 genes from the Metastasis group (High risk) and 4 genes from the Non-Metastasis group (Low risk) which totaled up to an 11 gene signature.

E2F had multiple probes and thus their functional properties were considered. A few subsets of the 11 hallmarks with the help of PubMed were considered to identify gene signatures.

E2F1, E2F2, and E2F3 were a family with functional similarities and E2F3 had the least significance among them. So it was ignored. Similarly E2F4 and E2F5 were another family with functional similarities and E2F5 is not as significant as E2F4. So it was ignored.

Thus E2F1, E2F2, and E2F4 were only included with the remaining 6 hallmarks to make a set of 9 hallmarks to find another signature. The genes that had interactions with all the 9 hallmarks were picked from both the Metastasis group (High risk) and the Non-Metastasis group (Low risk). There were 13 genes from the Metastasis group (High risk) and 8 genes from the Non-Metastasis group (Low risk) which totaled to a 21 gene signature.

Since E2F1 and E2F2 were a part of the same family, they had similar functionality and so they were considered one at a time to find more gene signatures.

When E2F1 and E2F4 were considered along with the remaining 6 hallmarks, there were a total of 8 hallmarks. All the genes that had interactions with the 8 hallmarks were picked from both the Metastasis group (High risk) and the Non-Metastasis group (Low risk). There were 18 genes from the Metastasis group (High risk) and 13 genes from the Non-Metastasis group (Low risk) which totaled to a 31 gene signature.

When E2F2 and E2F4 were considered along with the remaining 6 hallmarks, there were a total of 8 hallmarks. All the genes that had interactions with the 8 hallmarks were picked from both the Metastasis group (High risk) and the Non-Metastasis group (Low risk). There were 32 genes from the Metastasis group (High risk) and 19 genes from the Non-Metastasis group (Low risk) which totaled to a 51 gene signature. But there was one gene which was common to both the groups and thus the signature size becomes 50.

The number of genes identified from each group using different sets of hallmarks is shown in the Table 3-1 below.

**Table 3-1: Number of genes identified to have interactions with major cancer hallmarks in each prognosis group in each gene signature**

	genes from poor prognosis	genes from good prognosis
11 gene signature	7	4
21gene signature	13	8
31 gene signature	18	13
50 gene signature	32	19

Hence 4 signatures were discovered where each one had 11 genes, 21 genes, 31 genes, and 50 genes which are shown in Table 3-2 below. The 50 gene signature has a gene FLJ13059 that was extracted from both groups.

It can be seen that the 11 gene signature is a subset of all the remaining 3 signatures (21 gene, 31 gene, and 50 gene signatures) and the 21 gene signature is a subset of the remaining 2 signatures (31 gene and 50 gene signatures). This was because the set of hallmarks used for identifying the 31 gene signature and the 50 gene signature were subsets of the hallmarks used for identifying the other 2 signatures (11 gene and 21 gene signatures).

But the 31 gene signature and the 50 gene signature had a few unique genes each. Other than the 21 genes in common, they have just one more gene (ESM1) common to both the signatures.

**Table 3-2: All signatures obtained using different combinations of Hallmarks**

	50-GENE SIGNATURE	31-GENE SIGNATURE	21-GENE SIGNATURE	11-GENE SIGNATURE
1	ACTL6B	ACTL6B	ACTL6B	ACTL6B
2	ADAM3B	ADAM3B	ADAM3B	ADAM3B
3	FABP7	FABP7	FABP7	FABP7
4	PRR4	PRR4	PRR4	PRR4
5	RPL27A	RPL27A	RPL27A	RPL27A
6	SLC22A11	SLC22A11	SLC22A11	SLC22A11
7	TAC3	TAC3	TAC3	TAC3
8	215642_at	215642_at	215642_at	215642_at
9	DEFA5	DEFA5	DEFA5	DEFA5
10	PALM	PALM	PALM	PALM
11	SCGB2A2	SCGB2A2	SCGB2A2	SCGB2A2
12	BCDIN3	BCDIN3	BCDIN3	
13	GAL3ST1	GAL3ST1	GAL3ST1	
14	PCDHB3	PCDHB3	PCDHB3	
15	PCDHGA3	PCDHGA3	PCDHGA3	
16	PRKACA	PRKACA	PRKACA	
17	SAMD4B	SAMD4B	SAMD4B	
18	ACTR8	ACTR8	ACTR8	
19	CAP2	CAP2	CAP2	
20	CDKN2B	CDKN2B	CDKN2B	
21	TMEM135	TMEM135	TMEM135	
22	217363_x_at	C1orf68		
23	BRD2	dJ222E13.2		
24	C9	LOR		
25	COG5	SSFA2		
26	DDB1	TEX11		
27	DKFZP586P0123	217470_at		
28	ESM1	DAG1		
29	FLJ13059	ESM1		
30	GABRA1	H2AFB3		
31	HSPA2	TM4SF20		
32	KRT81			
33	MUC8			
34	PEX5L			
35	PPP1R2P9			
36	PRKAA1			
37	SUPT6H			
38	TPSD1			
39	TRIM9			
40	VPS35			
41	ATP6V0B			
42	CHD6			
43	DUSP21			
44	ELL			
45	KIAA1446			
46	SCN8A			
47	SLC26A1			
48	SPINK5			
49	STT3A			
50	TSPAN2			

### **3.3. Survival Analysis**

To find the most significant signature from the obtained prognostic signatures, survival analysis was performed on the four signatures. Survival Analysis was done using techniques which include Time dependant ROC analysis, Random Test, and Cox proportional hazard model. Cox model uses Kaplan-Meier plots and log-rank tests to identify the best signature.

#### **3.3.1 Time dependant ROC analysis and Random Test**

ROC curve stands for Receiver Operating Characteristic curve. It is a plot between the sensitivity and the (1-specificity) as time is varied. It can also be considered as the plot between the True Positive Rate and the False Positive Rate. That is it is used to describe the tradeoff between the hit rates and the false alarm rates. The higher the area under curve (AUC) values, the better the signature. The AUCs are calculated using R.

In Random test, genes are picked randomly from the entire set of genes. The number of genes picked must equal the number of genes in the signature that is being validated. The AUC of the signature genes is compared with the AUC of the picked genes. Similarly, a large number of randomly picked signatures are compared with the signature that is being validated. The performance of the identified gene signature must be significant when compared to the other randomly picked signatures. The lower the p-value from the random test is, the better is the signature.

In the datasets, when there were duplicate probes for genes, the probe that resulted in the most significant (least) p-value when fitted in to Cox model was considered the best probe and was used for time dependant ROC analysis of the entire signature. Thus time dependant ROC analysis [16] was done on the Training dataset (UM+HLM) and on both the Test datasets (DFCI



and MSK). The function “coxph” in R was used for the calculation of p-values to find the most significant probe from duplicate probes.

The Tables 3-3, 3-5, 3-7 show the AUC values over years starting from 1 to 9 and for all the four identified signatures in training dataset, DFCI test set and MSK test set respectively. The Tables 3-4, 3-6, 3-8 are the p-values of the corresponding signature obtained when compared to the randomly picked signatures from random test for training dataset, DFCI test set and MSK test set respectively.

**Table 3-3: AUC’s of training set (256 samples) obtained when best probes among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
1-year	0.684292	0.73329	0.777634	0.812282
2-years	0.648464	0.683333	0.729109	0.722581
3-years	0.661654	0.684145	0.711605	0.730173
4-years	0.642624	0.675829	0.704544	0.715059
5-years	0.640428	0.687573	0.707053	0.728794
6-years	0.64902	0.69424	0.703125	0.749081
7-years	0.644137	0.696312	0.711665	0.749984
8-years	0.628796	0.68712	0.699266	0.743927
9-years	0.618716	0.680933	0.693261	0.744874

**Table 3-4: p-values from Random test of training set (256 samples) obtained when best probes among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
3yr random	0.05	0.11	0.12	0.35
5yr random	0.32	0.26	0.38	0.78

As per the values shown in the Tables 3-3 above for Training dataset, ROC values were good for the 50 gene dataset. But according to Table 3-4, the Random test values were bad. The Random test values from Table 3-4 of the 11 gene and the 31 gene signatures showed better results. So a

tradeoff between the AUC values and the Random Test values points to the 31 gene signature which is nearly good in both the validations.

**Table 3-5: AUC's of DFCI test set (82 samples) obtained when best probes among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
1-year	0.610755	0.682458	0.773368	
2-years	0.660634	0.716742	0.80724	
3-years	0.712601	0.778187	0.847458	Error in fitter
4-years	0.706731	0.802885	0.866071	Ran out of iterations
5-years	0.719577	0.789021	0.831349	and did not converge
6-years	0.7225	0.780625	0.839375	
7-years	0.731618	0.77451	0.834559	
8-years	0.728875	0.774468	0.828571	
9-years	0.728875	0.774468	0.828571	

**Table 3-6: p-values from Random test of DFCI test set (82 samples) obtained when best probe among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
3 yr random	0.51	0.45	0.61	Error in fitter
5yr random	0.49	0.67	0.7	Ran out of iterations

From the Table 3-5 above for DFCI dataset, it was seen that the 50 gene signature had errors to fit in the model and hence was ruled out. Since the Random test values from Table 3-6 were not good for any signature in this dataset, only the AUC values were considered to decide that the 31 gene signature was better than the other signatures.

**Table 3-7: AUC's of MSK test set (104 samples) obtained when best probes among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
1-year	0.778878	0.933993	0.980198	
2-years	0.720284	0.78876	0.861757	
3-years	0.764359	0.828234	0.880837	Error in fitter
4-years	0.742069	0.816092	0.848736	Ran out of iterations
5-years	0.72479	0.833613	0.85	and did not converge
6-years	0.740196	0.841503	0.851307	
7-years	0.745865	0.842275	0.84994	

8-years	0.745865	0.842275	0.84994	
9-years	0.745865	0.842275	0.84994	

**Table 3-8: p-values from Random test of MSK test set (104 samples) obtained when best probes among duplicates were considered**

	11-gene	21-gene	31-gene	50-gene
3yr random	0.23	0.19	0.23	Error in fitter
5yr random	0.35	0.08	0.32	Ran out of iterations

From the values from the Table 3-7 above for MSK dataset, it was seen that the AUC values show in favor of the 31 gene signature where as the Random test values from Table 3-8 were in favor of the 21 gene signature.

Thus considering all the three datasets, Time dependant ROC values and Random test together were more in favor of the 31 gene signature over the other signatures.

### **3.3.2 Cox proportional hazards model on 11, 21 and 31 Gene Signatures**

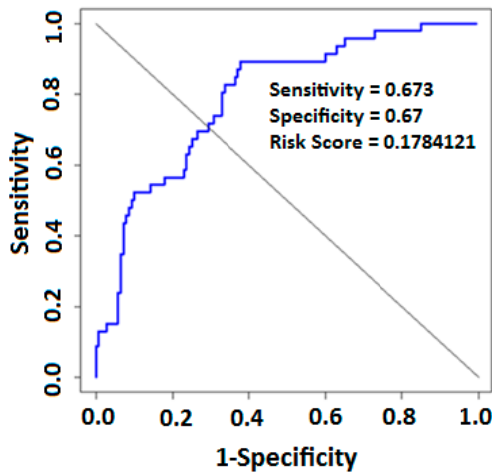
For the COX modeling, the training dataset is first fit in to the Cox model and then the cutoff values obtained by the fitting are applied on the test datasets. Three different cutoffs from training dataset were used which are mean, median and the nearest point.

#### **Mean or Median as cutoff:**

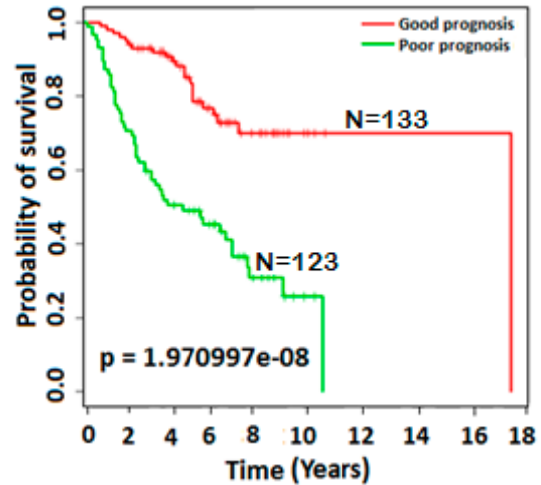
The means/medians of the samples of the training dataset were calculated and were applied as the cutoffs for the test datasets. Kaplan Meier plots were plotted to see if the stratification was significant. Only three signatures (11-genes, 21-genes and 31-genes) were considered for evaluation in Cox model. The fourth signature was ignored as it gave errors in the previous models. The results are shown in Tables 3-9, 3-10 and 3-11 for 11-gene, 21-gene, and 31-gene

signatures respectively. The Kaplan Meier plots were not displayed for mean/median as cutoff, as they were better for the nearest point cutoff which will be discussed below.

**Nearest point as cutoff:**



**Figure 3-3: Finding Nearest point as cutoff for Training data for 31 genes**



**Figure 3-4: KM plot of Training data with nearest point cutoff for 31 genes**

In this process, the training dataset for each signature is fit in the Cox model for the predict time equaling 3 years. The time dependant ROC curve is plotted. Then the cutoff point has to be chosen. To choose the cutoff point, the distance from the point sensitivity=1 to each and every point on the time dependant ROC curve plotted before is calculated. The cutoff point would be none other than the point whose distance is the minimum. In other words, it is the point on the curve nearest to the left top corner of the plot. The point would always be the point of intersection of the time dependant ROC curve and the diagonal (not passing through origin). The plot is shown in Figure 3-3. Kaplan Meier plots are drawn for the 31 gene signature to see if the stratification of the data was significant or not as shown in Figure 3-4. Since the stratification from KM plot was very good and the log-rank p-value of the dataset was very significant, this

cutoff point from training set was used to fit the test datasets in the Cox model. Kaplan Meier plots of the outputs were drawn as shown in Figure 3-5 and Figure 3-6.

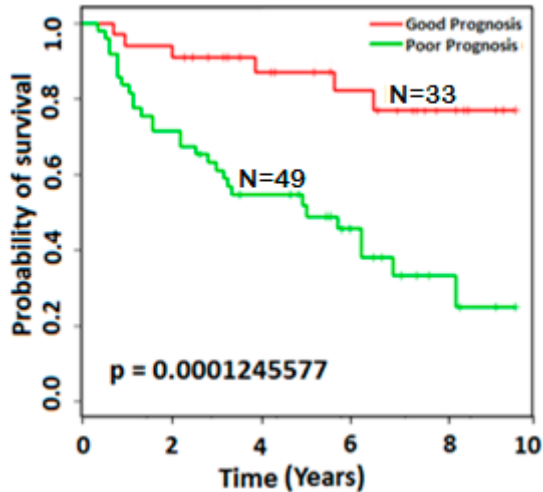


Figure 3-5: KM plot of DFCI data with nearest point cutoff of training data for 31 gene signature

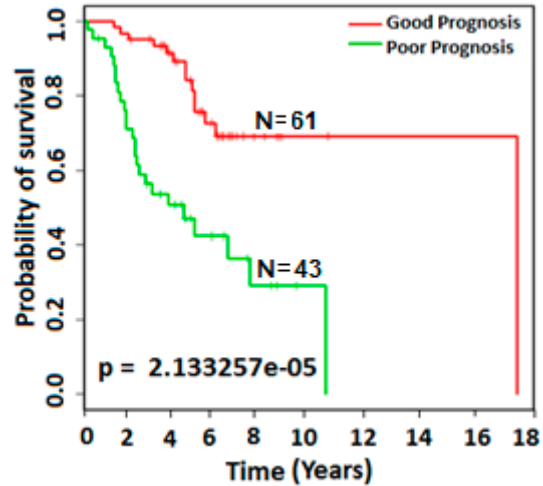


Figure 3-6: KM plot of MSK data with nearest point cutoff of training data for 31 gene signature

Table 3-9: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 11-gene signature

	train cutoff	train output	MSK test output	DFCI test output
11-gene signature	median	Train p-value 0.02989034 Train cutoff -1.521898	Test p-value 0.08099651 Test cutoff -1.521898	Test p-value 0.2264746 Test cutoff -1.521898
	mean	Train p-value 0.007076216 Train cutoff -1.502247	Test p-value 0.03688842 Test cutoff -1.502247	Test p-value 0.109008 Test cutoff -1.502247
	nearest point	Train p-value 0.007076216 Train cutoff -1.501426	Test p-value 0.03688842 Test cutoff -1.501426	Test p-value 0.109008 Test cutoff -1.501426

From the Table 3-9 above, it can be seen that the log-rank p-values for 11 gene signature for MSK dataset were significant but they were not significant for the DFCI dataset for any kind of cutoff values.

**Table 3-10: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 21-gene signature**

	train cutoff	train output	MSK test output	DFCI test output
21-gene signature	median	Train p-value 1.86E-05 Train cutoff -0.7944426	Test p-value 3.99E-05 Test cutoff -0.7944426	Test p-value 0.04601307 Test cutoff -0.7944426
	mean	Train p-value 3.51E-05 Train cutoff -0.8304135	Test p-value 5.57E-05 Test cutoff -0.8304135	Test p-value 0.06825427 Test cutoff -0.8304135
	nearest point	Train p-value 1.26E-07 Train cutoff -0.63574	Test p-value 1.58E-07 Test cutoff -0.63574	Test p-value 0.01306106 Test cutoff -0.63574

From the Table 3-10 above, it can be seen that the log-rank p-values for 21 gene signature were significant, except when the mean was used as a cutoff for DFCI dataset.

**Table 3-11: Cox model outputs for various cutoffs of training dataset applied on both DFCI and MSK test datasets for 31-gene signature**

	train cutoff	train output	MSK test output	DFCI test output
31-gene signature	median	Train p-value 1.35E-08 Train cutoff 0.1062929	Test p-value 4.59E-05 Test cutoff 0.1062929	Test p-value 0.000148722 Test cutoff 0.1062929
	mean	Train p-value 1.16E-07 Train cutoff 0.1238266	Test p-value 2.13E-05 Test cutoff 0.1238266	Test p-value 0.000607811 Test cutoff 0.1238266
	nearest point	Train p-value 1.97E-08 Train cutoff 0.1784121	Test p-value 2.13E-05 Test cutoff 0.1784121	Test p-value 0.000124558 Test cutoff 0.1784121

From the Table 3-11 above, it can be seen that the log-rank p-values for 31 gene signature were very significant in all the datasets.

Since all the above methods showed good results for 31 gene signature, it was considered to be further validated. Though the 21 gene signature had almost good results, 31 gene signature was considered as it included all the 21 genes from the 21 gene signature.

Since 31 gene signature performed well in the above validation techniques, we tried to evaluate the performance of just the Stage I patients predicted from the above model.

There were 157 patients in the Training dataset who belonged to Stage I and among them 83 patients were predicted as Good Prognosis and 74 patients were predicted as Poor Prognosis. The log-rank p-value and the KM plot shown in Figures 3-7 below show that Stage I is very significant in the Training dataset.

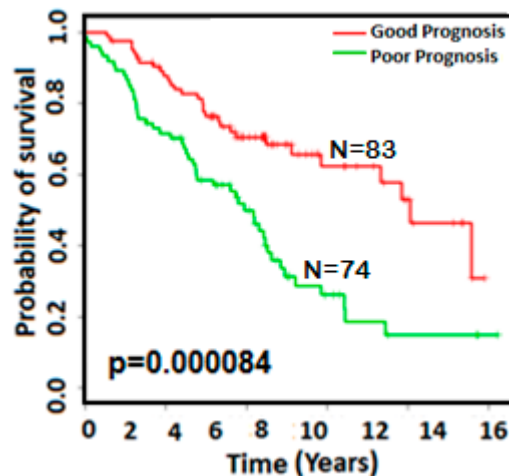


Figure 3-7: KM plot of Training data for Stage I patients

There were 56 patients in DFCI dataset who belonged to Stage I and among them 24 patients were predicted as Good Prognosis and 32 patients were predicted as Poor Prognosis. There were

63 patients in MSK dataset who belonged to Stage I and among them 41 patients were predicted as Good Prognosis and 22 patients were predicted as Poor Prognosis.

Thus it can be seen from the Figures 3-8 and 3-9 below that the stratification for Stage I patients was very significant and that the log-rank p-values were very significant.

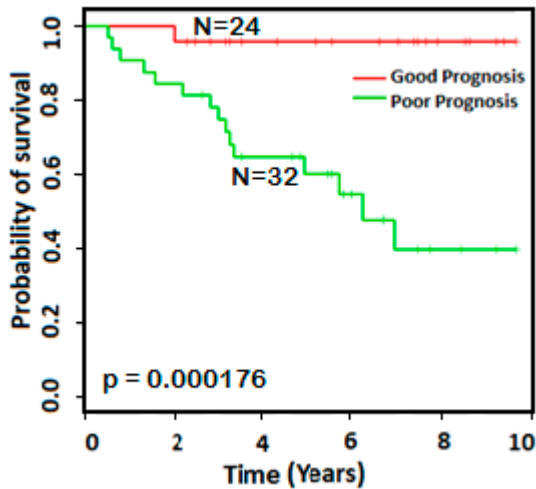


Figure 3-8: KM plot of DFCI data for Stage I patients

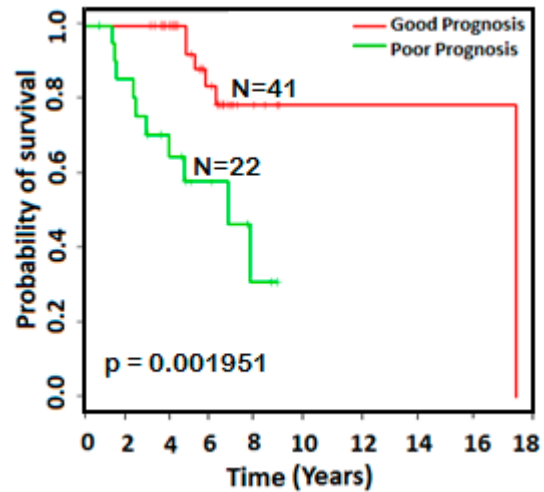


Figure 3-9: KM plot of MSK data for Stage I patients

Then the Stage I was further split in to Stage IA and Stage IB patients and the same analysis was performed.

There were 76 patients in the training dataset (UM+HLM) who belonged to Stage IA and among them 48 patients were predicted as Good Prognosis and 28 patients were predicted as Poor Prognosis. Thus the training dataset (UM+HLM) was significant for Stage IA group according to the log-rank p-value and KM plot as shown in Figure 3-10 below.



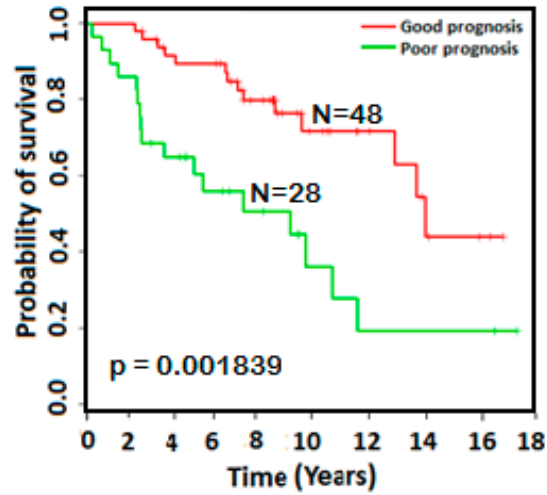


Figure 3-10: KM plot of Training data for Stage IA patients

There were 11 patients in DFCI dataset who belonged to Stage IA and among them 5 patients were predicted as Good Prognosis and 6 patients were predicted as Poor Prognosis. There were 27 patients in MSK dataset who belonged to Stage IA and among them 18 patients were predicted as Good Prognosis and 9 patients were predicted as Poor Prognosis.

According to the KM plots and log-rank p-values shown in Figures 3-11 and 3-12, Stage IA of the Test sets (DFCI and MSK) was not as significant as Stage I, which might be because of the fewer number of patients belonging to Stage IA in the analysis.

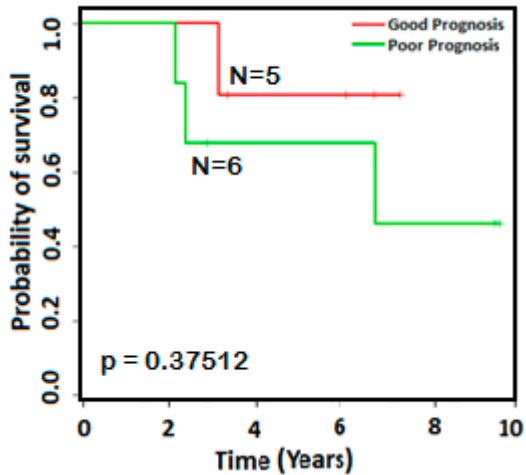


Figure 3-11: KM plot of DFCI data for Stage IA patients

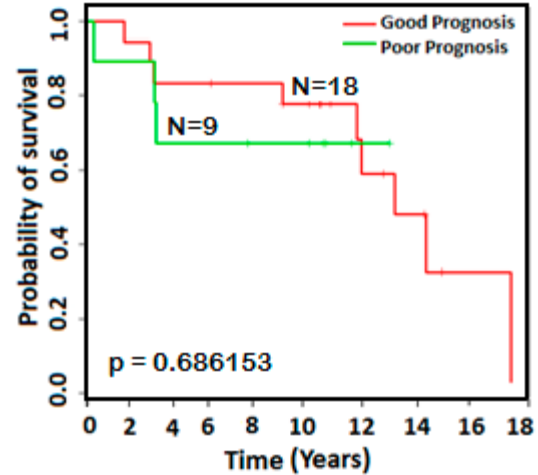


Figure 3-12: KM plot of MSK data for Stage IA patients

There were 81 patients in the training dataset (UM+HLM) who belonged to Stage IB and among them 35 patients were predicted as Good Prognosis and 46 patients were predicted as Poor Prognosis as shown in the KM plot in Figure 3-13 below.

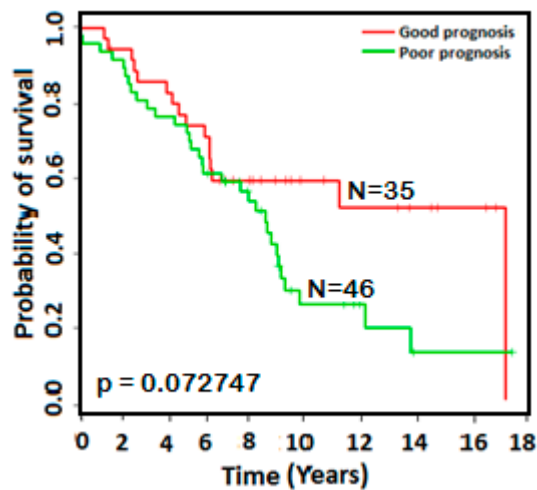


Figure 3-13: KM plot of Training data for Stage IB patients

There were 45 patients in DFCI dataset who belonged to Stage IB and among them 19 patients were predicted as Good Prognosis and 26 patients were predicted as Poor Prognosis. There were

36 patients in MSK dataset who belonged to Stage IB and among them 23 patients were predicted as Good Prognosis and 13 patients were predicted as Poor Prognosis.

Thus Stage IB in DFCI and MSK datasets was significant from the log-rank p-values and KM plots in the Figures 3-14 and 3-15 shown below.

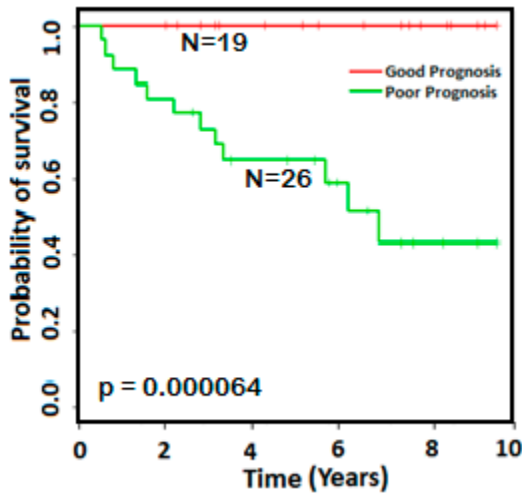


Figure 3-14: KM plot of DFCI data for Stage IB patients

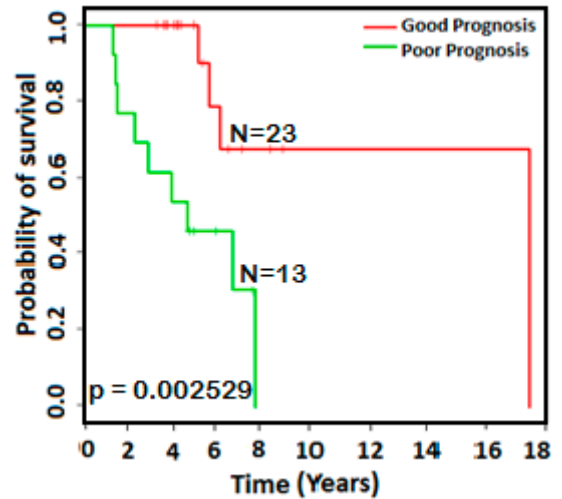


Figure 3-15: KM plot of MSK data for Stage IB patients

### 3.4. Summary

This chapter provided a flow chart of the methodology used and discussed how the gene signatures were identified using implication networks. The details of the datasets used were mentioned and all the procedures used and the results obtained by the methodology were also summarized. The major cancer hallmarks which were used to identify signatures were described and all the gene signatures identified were given. Survival analysis results were provided for all the datasets. Thus the 31 gene signature was considered to be the most significant from all the analysis done and it will be considered for further evaluation.

## **4 Prognostic Validation, Clinical Evaluation & Topological Validation**

### **4.1 Introduction**

In the earlier chapter, details were provided on the identification of gene signatures. This chapter discusses about the validation techniques that were used and all the processing that was done on the datasets. Details about the gene signature were provided. Prognostic validation performed done using techniques like Concordance probability estimates and Gene set enrichment analysis. Clinical evaluation was conducted using Multivariate COX proportional hazards model. Topological validation was done by comparing the interactions from implication networks with interactions from Bayesian networks built using Tetrad IV. Various web based tools were also used to confirm the presence and the significance of the interactions from implication networks. All the results that were obtained from Prognostic validation, Clinical evaluation, and Topological validation are provided.

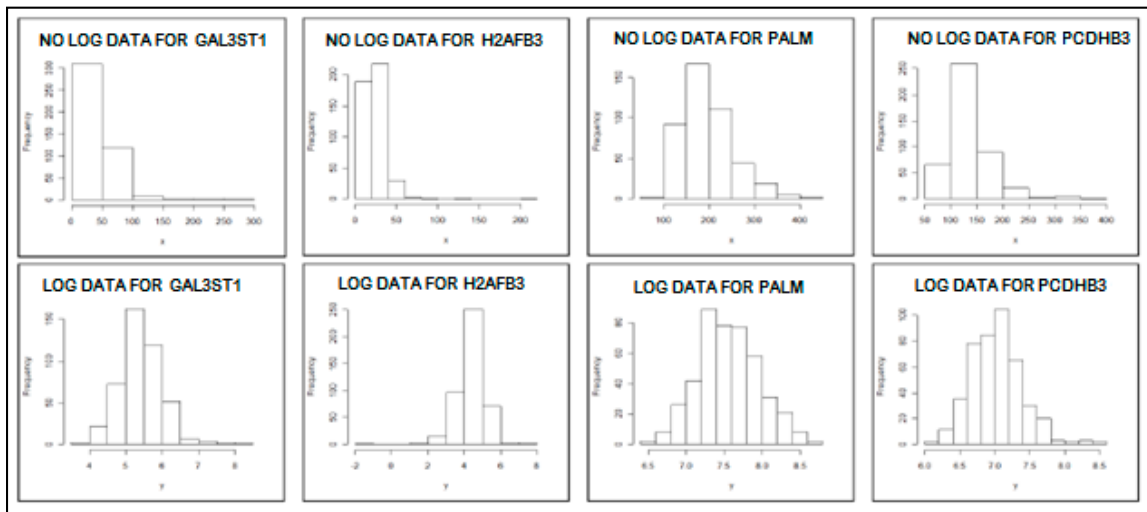
### **4.2 Gene Signature Details and Differentially Expressed genes**

Since the 31 gene signature was considered to be the most prognostic signature, the details of the 31 genes were provided. The details include the chromosome locations, molecular functions and classifications that have been confirmed by Dr. Yong Qian from The National Institute for Occupational Safety and Health (NIOSH). These details are shown in the Table 4-1 below.

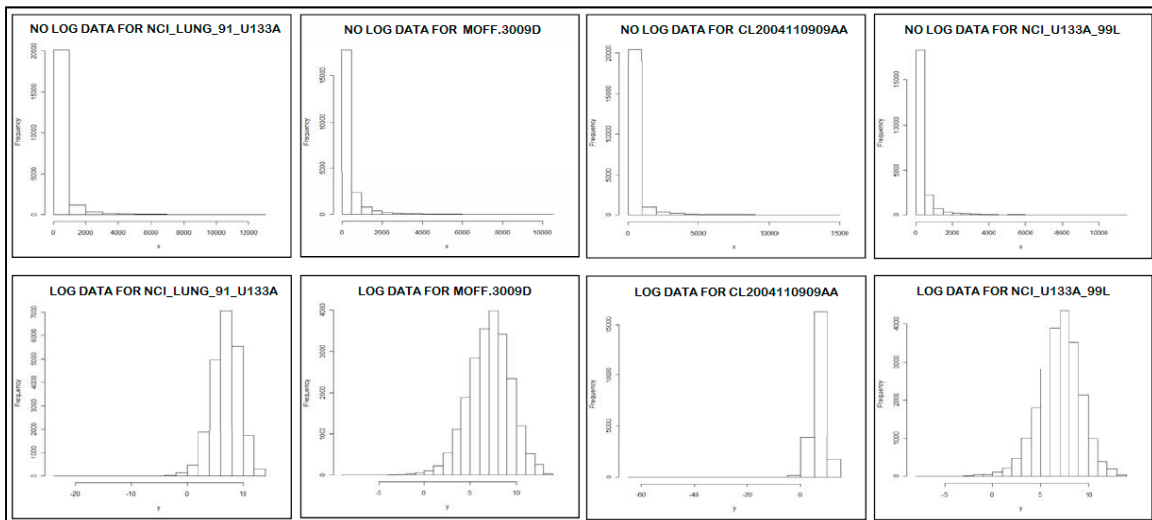
**Table 4-1: Details of the 31 gene signature that have been confirmed by Dr. Yong Qian from NIOSH**

Name	PROBESET ID (of the best probe)	Chromosome location	Molecular function	Classification
Unknown	215642_at	10q23.33		
Unknown	217470_at	4q35.2		
ACTL6B	206014_at	7q22	Vesicular transport, spindle orientation, nuclear migration and chromatin remodeling	Structure
ACTR8	218658_s_at	3		Structure
ADAM3B	217237_at	16q12.1		
BCDIN3	219798_s_at	7q22.1	S-adenosyl-L-methionine-dependent methyltransferase	Metabolism
C1orf68	217087_at	1q21.3		
CAP2	212554_at	6p22.3		
CDKN2B	207530_s_at	9p21	Cell growth regulator	Oncogene
DAG1	205417_s_at	3p21	Link the cytoskeleton to the extracellular matrix	Structure
DEFA5	207529_at	8pter-p21	Host defense	Immunity
dJ222E13.2	214828_s_at	22q13.2		
ESM1	208394_x_at	5q11.2	Lung endothelial cell-leukocyte interactions	Signaling transduction
FABP7	216192_at	6q22-q23	Fatty acid uptake, transport, and metabolism	Metabolism
GAL3ST1	205670_at	22q12.2	Sulfotransferase activity	Metabolism
H2AFB3	214412_at	Xq28	A protein component of histone	Structure
LOR	207720_at	1q21	A major protein component of the cornified cell envelope	Structure
PALM	203859_s_at	19p13.3	control of cell shape	Structure
PCDHB3	221410_x_at	5q31	Establishment and maintenance of specific neuronal connections in the brain	Structure
PCDHGA3	209478_at	5q31	Establishment and maintenance of specific neuronal connections in the brain	Structure
PRKACA	216234_s_at	19p13.1	Protein kinase	Signaling transduction
PRR4	204919_at	12p13	Protection in eye	Immunity
RPL27A	203034_s_at	11p15	A component of ribosome	Metabolism
SAMD4B	220457_at	19q13.2		
SCGB2A2	206378_at	11q13		
SLC22A11	220100_at	11q13.1	Mediates saturable uptake of estrone sulfate, dehydroepiandrosterone sulfate and related compounds	Metabolism
SSFA2	202506_at	2q31.3		
TAC3	219992_at	12q13-q21	Vasodilators and secretagogues	Signaling transduction
TEX11	221259_s_at	Xq13.1		
TM4SF20	220639_at	2q36.3		
TMEM135	222209_s_at	11q14.2		

To find the differentially expressed genes from the 31 gene signature, T-tests and fold changes were calculated. To perform the Fold change analyses, data should be distributed normally. Since the data we have been using is not normally distributed, we log transform the data to change it to the required form. The histograms of a few genes over all the samples and histograms of a few samples over all the genes were plotted as shown below in the Figures 4-1 and 4-2.



**Figure 4-1: Histograms for 4 genes over all the 442 samples of data showing that the log transformed data is less skewed than data which was not log transformed**



**Figure 4-2: Histograms for 4 samples, each from one dataset over all the 22215 genes of data showing that the log transformed data is less skewed than data which was not log transformed**

To find the differentially expressed genes, T-tests and fold changes analyses were performed. Their results for the 31 genes are shown below in Table 4-2 and Table 4-3. The data for all the 442 samples (UM+HLM+DFCI+MSK) was considered together. T-tests for the 31 genes with respect to the different predictive factors such as Stage, Tumor Differentiation and Lymph node metastases are performed. The genes that were significant in T-test ( $\leq 0.05$ ) are shown with a star in the plots shown for fold changes further below.

There were three stages (1, 2 and 3) and calculations were conducted with respect to stage 1 samples. For analysis with Stage, T-tests for Stage 2 to Stage 1 had four significant genes and T-tests for Stage 3 to Stage 1 also had four significant genes. They are shown with a star in the Figure 4-3 below.

There were three kinds of tumor differentiation (well, poorly and moderate differentiated) and the calculations were performed with respect to well differentiated samples. For analysis with Tumor Differentiation, T-tests for Moderate differentiation to Well differentiation had six significant genes and T-tests for Poor Differentiation to Well differentiation had eight significant genes. They are shown with a star in the Figure 4-4 below.

There were two kinds of lymph node metastases (LN- and LN+) where calculation was done with respect to LN- samples. For analysis with Lymph node metastases, T-tests for lymph node positive to lymph node negative had no significant genes. All the results obtained for T-tests are shown below in Table 4-2.

**Table 4-2: T-test outputs for different predictors such as Stage (Stage-2 to Stage-1 and Stage-3 to Stage-1), Tumor differentiation (Moderate to Well and Poor to Well) and Lymph node metastases (LN+ to LN-) outputs for all the 31 genes in the signature**

Gene Symbols	T-test for stage 2 to 1	T-test for stage 3 to 1	T-test for Tumor Differentiation Moderate to Well	T-test for Tumor Differentiation Poor to Well	T-test for Lymph node metastases LN+ to LN-
215642_at	0.441814	0.145264	0.857738	0.738474	0.322506
217470_at	0.655542	0.338437	0.661369	0.131298	0.393442
ACTL6B	0.727608	0.607288	0.768965	0.173359	0.864637
ACTR8	0.033593	0.750247	0.551963	0.856646	0.055846
ADAM3B	0.938388	0.068125	0.013012	0.000868	0.288685
BCDIN3	0.858558	0.223081	0.289013	0.802365	0.648078
C1orf68	0.022491	0.700184	0.169695	0.012192	0.193197
CAP2	0.768851	0.457084	0.571463	0.000115	0.96312
CDKN2B	0.446325	0.945302	0.264274	0.894857	0.307397
DAG1	0.542864	0.775556	0.97073	0.350751	0.855356
DEFA5	0.259804	0.189907	0.160888	0.180041	0.729967
dJ222E13.2	0.595055	0.729298	0.15501	0.765353	0.262961
ESM1	0.039148	0.280311	0.000309	0.000389	0.050772
FABP7	0.838281	0.818263	0.713371	0.193622	0.929565
GAL3ST1	0.727634	0.01859	0.039144	0.002981	0.57002
H2AFB3	0.322762	0.528597	0.743628	0.448308	0.514133
LOR	0.469576	0.575702	0.987738	0.124489	0.998675
PALM	0.267144	0.685552	0.825739	0.884183	0.557783
PCDHB3	0.652335	0.085702	0.029541	0.067242	0.603537
PCDHGA3	0.354072	0.105162	0.32659	0.93396	0.470739
PRKACA	0.16995	0.224605	0.263827	0.023981	0.844798
PRR4	0.765668	0.02991	0.194364	0.109536	0.742261
RPL27A	0.553478	0.011909	0.079101	0.962865	0.378543
SAMD4B	0.877808	0.295199	0.405373	0.134323	0.935369
SCGB2A2	0.242397	0.298675	0.94019	0.769274	0.493917
SLC22A11	0.34981	0.770405	0.930257	0.986529	0.657845
SSFA2	0.024944	0.051833	0.366235	0.005074	0.330087
TAC3	0.258139	0.087507	0.903948	0.249307	0.743349
TEX11	0.164771	0.161557	0.006716	0.228346	0.695372
TM4SF20	0.858663	0.80129	0.654298	0.818728	0.993537
TMEM135	0.618155	0.003059	0.00016	0.039461	0.820871



Fold Changes for the 31 genes with respect to the different predictive factors such as Stage, Tumor Differentiation and Lymph node metastases are shown below in Figures 4-3, 4-4, and 4-5 respectively. The genes that have fold changes  $\geq 2$  are considered to be upregulated and the genes that have fold changes  $\leq 0.5$  are considered to be down regulated. There were no genes that were upregulated or down regulated according to Fold changes.

There were three stages (1, 2 and 3) and fold change calculations were conducted with respect to stage 1 samples which are shown in Figure 4-3. There were three kinds of tumor differentiation (well, poorly and moderate differentiated) and the calculations were performed with respect to well differentiated samples which are shown in Figure 4-4. There were two kinds of lymph node metastases (LN- and LN+) where calculation was done with respect to LN- samples as shown in Figure 4-5. Error bars showing the 95% confidence intervals are also shown in the figures.

**Table 4-3: Fold changes for different predictors such as Stage (Stage-2 to Stage-1 and Stage-3 to Stage-1), Tumor differentiation (Moderate to Well and Poor to Well) and Lymph node metastases (LN+ to LN-) outputs for all the 31 genes in the signature**

Gene Symbols	Fold Change for stage 2 to 1 = $\Delta 2/1$	Fold Change for stage 3 to 1 = $\Delta 3/1$	Fold Change for Tumor Differentiation Moderate to Well = $\Delta \text{Moderate/Well}$	Fold Change for Tumor Differentiation Poor to Well = $\Delta \text{Poor/Well}$	Fold Change for Lymph node metastases LN+ to LN- = $\Delta \text{LN+}/\text{LN-}$
215642_at	1.024592	0.864067	0.96636	1.168628	1.047483
217470_at	0.867161	1.641676	1.585661	0.622946	0.893683
ACTL6B	0.559918	1.620453	1.347857	0.609641	0.722991
ACTR8	0.378939	90.51783	42.39918	0.113063	0.236895
ADAM3B	0.559942	6.168879	2.771941	1.491585	0.854981
BCDIN3	4.87E-05	6.66E+12	84896277	2.63E-09	0.000392
C1orf68	0.023504	317.101	51.31794	0.033108	0.141262
CAP2	0.369338	72610.67	58115.3	37.67143	0.757832
CDKN2B	0.260977	2.634748	2.371027	0.473173	0.43117
DAG1	0.000011	9.95E+23	5.51E+18	2.51E-09	2.44E-05
DEFA5	6.41E-19	282.4736	135.1423	6.94E-11	3.8E-13
dJ222E13.2	0.030751	230.7319	13.75698	0.023975	0.032192
ESM1	0.000932	254493.4	9091.178	0.023702	0.034733
FABP7	6.18482	1673.235	0.253757	0.017357	4.768529
GAL3ST1	0.066054	228.7593	19.25724	2.281099	0.28678
H2AFB3	0.473328	4.360512	7.784657	0.856106	0.494805
LOR	0.888696	3.77847	1.710162	0.877704	1.01584
PALM	0.237886	7.77E+08	1544293	0.000466	0.101454
PCDHB3	0.028522	450103.8	150.0798	3.77E-05	0.036493
PCDHGA3	0.820507	4.22E+18	502.1335	1.06E-11	0.207464
PRKACA	0.061531	131.563	18.31498	0.017633	0.112883
PRR4	2.905829	1447.149	143.6872	0.054177	2.557132
RPL27A	1.21E-82	3.3E+171	2.6E+124	2.66E-58	1.99E-69
SAMD4B	0.500379	7.762561	0.908299	0.304523	1.188885
SCGB2A2	0.95827	1.18309	0.922625	0.512091	0.908788
SLC22A11	0.759165	1.032998	2.002644	0.775557	0.717989
SSFA2	24945.98	2.12E+16	5.86E+24	1.584037	0.000148
TAC3	5.948372	2663.06	37.05733	0.0004	3.777423
TEX11	0.093164	233.0746	64.09065	0.249336	0.221878
TM4SF20	2.79735	14.56566	2217.304	408.6983	2.20726
TMEM135	0.001094	118.9389	0.02566	2.61E-12	5.41E-05

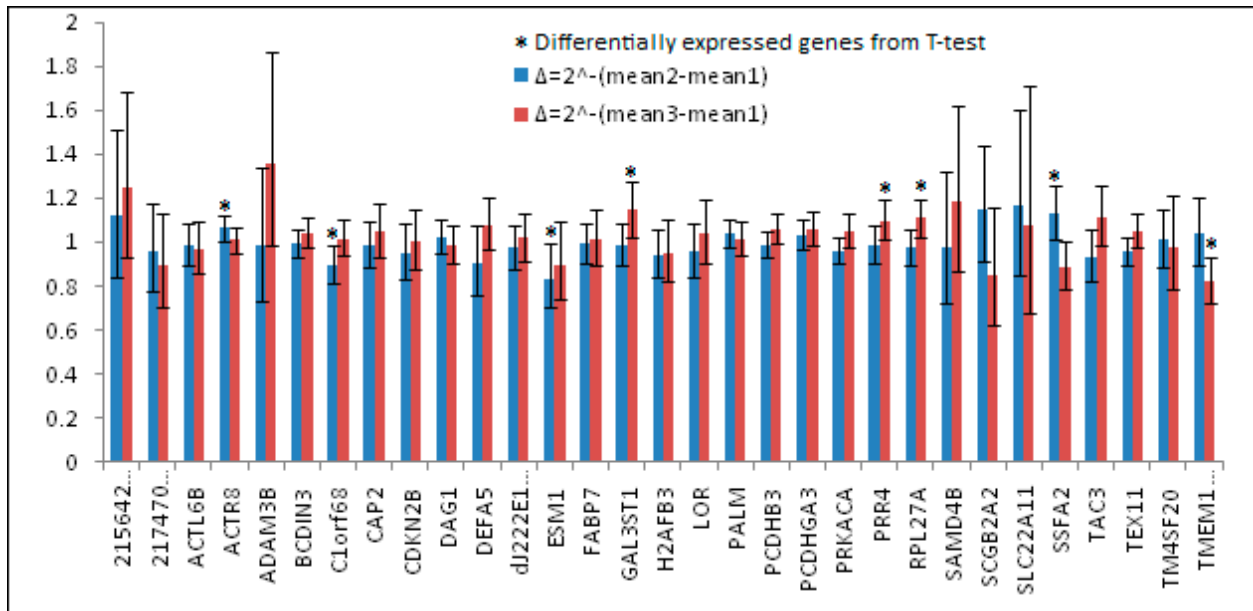


Figure 4-3: Fold changes of the 31 genes for Stage, where blue color bars represent fold change of stage 2 w.r.t. stage 1 and red color bars represent fold change of stage 3 w.r.t. stage 1 and genes with stars on the top represent the significant genes from T-test with  $p \leq 0.05$ .

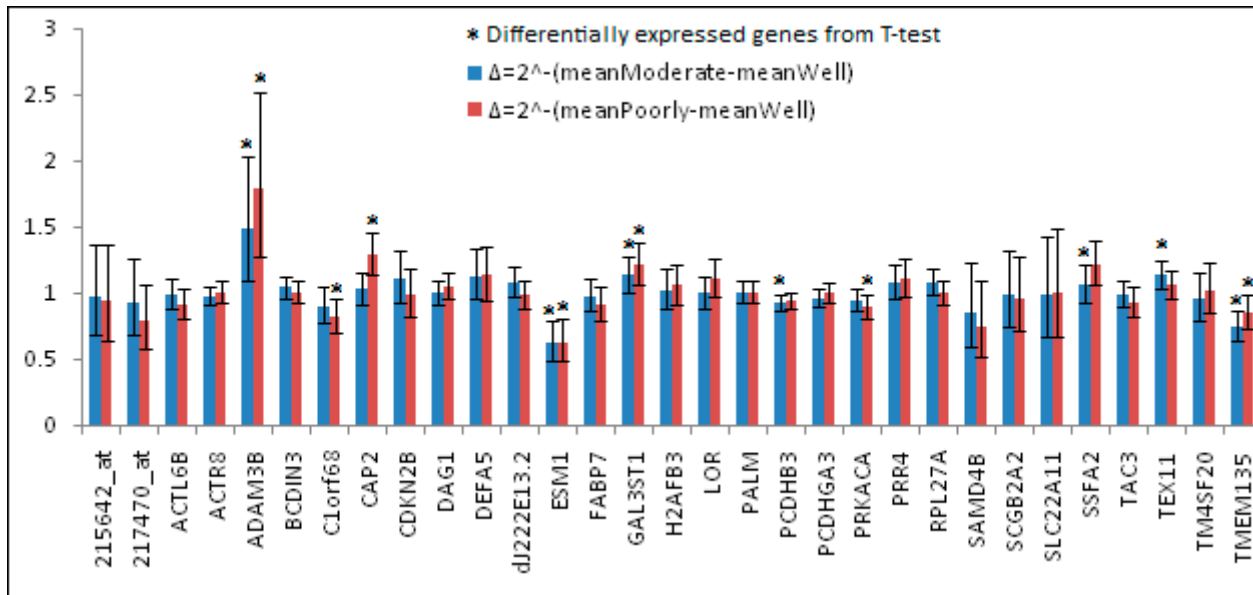


Figure 4-4: Fold changes of the 31 genes for Tumor Differentiation, where blue color bars represent fold change of Moderate differentiation w.r.t. Well differentiation and red color bars represent fold change of Poor differentiation w.r.t. Well differentiation and genes with stars on the top represent the significant genes from T-test with  $p \leq 0.05$ .

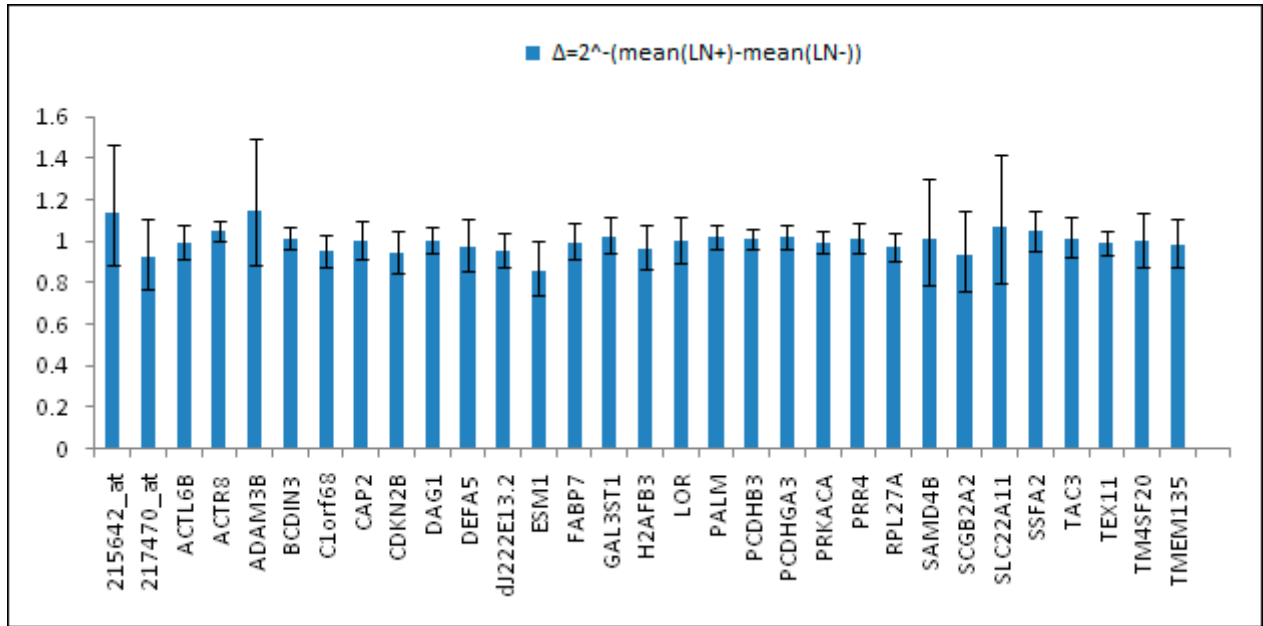


Figure 4-5: Fold changes of the 31 genes for Lymph node metastases, where blue color bars represent fold change of lymph node positive w.r.t. lymph node negative.

### 4.3 Prognostic Validation

Prognostic Validation is done using metrics such as Overall Accuracy, Concordance Probability Estimate (CPE), and Gene Set Enrichment Analysis (GSEA). Our model is also validated using other methods in Weka. Now only the 31 gene signature is considered for validation.

#### 4.3.1 Overall Accuracy:

##### Training data:

Overall Accuracy of the Training data for 31 gene signature was calculated with 3 years as a cutoff. Poor prognosis corresponds to samples that were dead by the end of the cut off period (in actual data) or were predicted as high risk through the model (in predicted data). Similarly, good prognosis corresponds to the samples that had their status to be living by the end of cutoff period (in actual data) or were predicted as low risk through the model (in predicted data). The contingency table was formed from the comparison of actual data with the predicted data. The 9

censored cases were ignored. There were 67 cases in Training data which belonged to Poor prognosis in both actual data as well as the predicted data which constituted the True Positives. Similarly there were 88 cases which belonged to Good prognosis in both actual and predicted data which constituted the True Negatives. Similarly False Positives and False Negatives were determined for the Training data which are shown below in the Table 4-4 along with the sensitivity, specificity and overall accuracy. The contingency table for the actual data versus the predicted data is shown below.

		Predicted data	
		Poor	Good
Actual data	Poor	TP=67	FP=64
	Good	FN=28	TN=88

**Table 4-4: Sensitivity, Specificity and Overall Accuracy of Training data calculated from contingency table**

Training	Actual data	Predicted data	Sensitivity	Specificity	Overall Accuracy
Poor prognosis	95	133	TP / (TP + FN) = 0.705263	TN / (FP + TN) = 0.578947	(TP+TN)/(TP+FP+FN+TN) =0.62753
Good prognosis	152	123			

**DFCI data:**

Overall Accuracy of the DFCI data for 31 gene signature was calculated with 3 years as a cutoff. Poor prognosis corresponds to samples that were dead by the end of the cut off period (in actual data) or were predicted as high risk through the model (in predicted data). Similarly, good prognosis corresponds to the samples that had their status to be living by the end of cutoff period (in actual data) or were predicted as low risk through the model (in predicted data). The contingency table was formed from the comparison of actual data with the predicted data. The 5 censored cases were ignored. There were 19 cases in Training data which belonged to Poor

prognosis in both actual data as well as the predicted data which constituted the True Positives. Similarly there were 26 cases which belonged to Good prognosis in both actual and predicted data which constituted the True Negatives. Similarly False Positives and False Negatives were determined for the DFCI data which are shown below in the Table 4-5 along with the sensitivity, specificity and overall accuracy. The contingency table for the actual data versus the predicted data is shown below.

		Predicted data	
		Poor	Good
Actual data	Poor	TP=19	FP=29
	Good	FN=3	TN=26

**Table 4-5: Sensitivity, Specificity and Overall Accuracy of DFCI data calculated from contingency table**

DFCI	Actual data	Predicted data	Sensitivity	Specificity	Overall Accuracy
Poor prognosis	22	49	$TP / (TP + FN)$ $= 0.863636$	$TN / (FP + TN)$ $= 0.472727$	$(TP+TN)/(TP+FP+FN+TN)$ $=0.584416$
Good prognosis	55	33			

**MSK data:**

Overall Accuracy of the MSK data for 31 gene signature was calculated with 3 years as a cutoff. Poor prognosis corresponds to samples that were dead by the end of the cut off period (in actual data) or were predicted as high risk through the model (in predicted data). Similarly, good prognosis corresponds to the samples that had their status to be living by the end of cutoff period (in actual data) or were predicted as low risk through the model (in predicted data). The contingency table was formed from the comparison of actual data with the predicted data. The 10 censored cases were ignored. There were 19 cases in Training data which belonged to Poor prognosis in both actual data as well as the predicted data which constituted the True Positives. Similarly there were 52 cases which belonged to Good prognosis in both actual and predicted

data which constituted the True Negatives. Similarly False Positives and False Negatives were determined for the MSK data which are shown below in the Table 4-6 along with the sensitivity, specificity and overall accuracy. The contingency table for the actual data versus the predicted data is shown below.

		Predicted data	
		Poor	Good
Actual data	Poor	TP=19	FP=19
	Good	FN=4	TN=52

**Table 4-6: Sensitivity, Specificity and Overall Accuracy of MSK data calculated from contingency table**

MSK	Actual data	Predicted data	Sensitivity	Specificity	Overall Accuracy
Poor prognosis	23	43	TP / (TP + FN) = 0.826087	TN / (FP + TN) = 0.732394	(TP+TN)/(TP+FP+FN+TN) =0.755319
Good prognosis	71	61			

### 4.3.2 Concordance Probability Estimate (CPE)

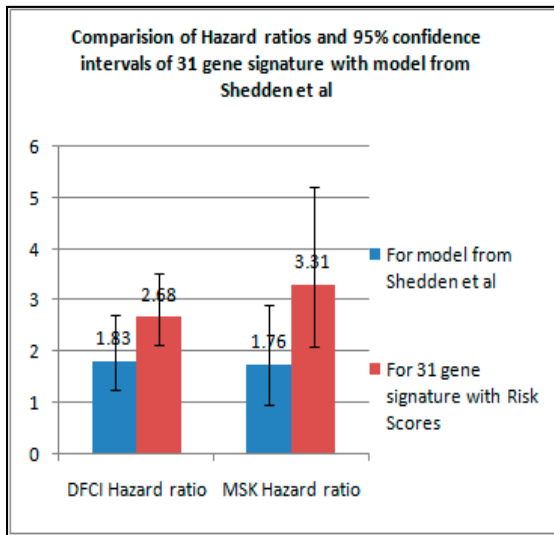
CPE [29] was calculated for the 31 gene signature using the function “phcpe” from R for both the test datasets DFCI and MSK. The outputs of the function were compared with the model from Shedden et al [20] as shown in Table 4-7.

**Table 4-7: Comparison of 31 gene signature with model from Shedden et al [20] on both the Test datasets where log-rank p-values, hazard ratios, and confidence intervals were obtained from the CPE package which use the risk scores of the entire signature as input**

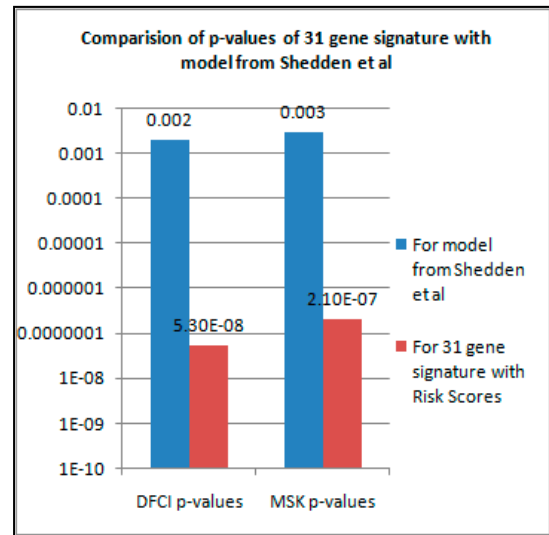
DFCI dataset	Hazard ratio with 95% CI	Log-rank p-value	CPE
For model from Shedden et al [20]	1.83 [1.24, 2.70]	0.002	0.63
For 31 gene signature with Risk Scores	2.68 (1.88, 3.82]	5.30E-08	0.71
MSK dataset	Hazard ratio with 95% CI	Log-rank p-value	CPE
For model from Shedden et al [20]	1.76 [1.20, 2.60]	0.003	0.62
For 31 gene signature with Risk Scores	3.31 [2.11, 5.2]	2.10E-07	0.70

The hazard ratios for both the test datasets were higher for the 31 gene signature which shows that the signature has strong capability of estimating the risk. The hazard ratios and the 95%

confidence intervals are shown by error bars in the Figure 4-6. The log-rank p-values of the test datasets are much lower for the 31 gene signature when compared to the results from the model from Shedden et al [20] as shown in Figure 4-7 which shows that they are highly significant. The CPE values are supposed to be higher than 0.5 and as high as possible. Figure 4-8 shows that the 31 gene signature had much higher CPE values than the model from Shedden et al [20]. This proves that the 31 gene signature has better performance in terms of CPE, hazard ratios, and log-rank p-values when compared to the model from Shedden et al [20].

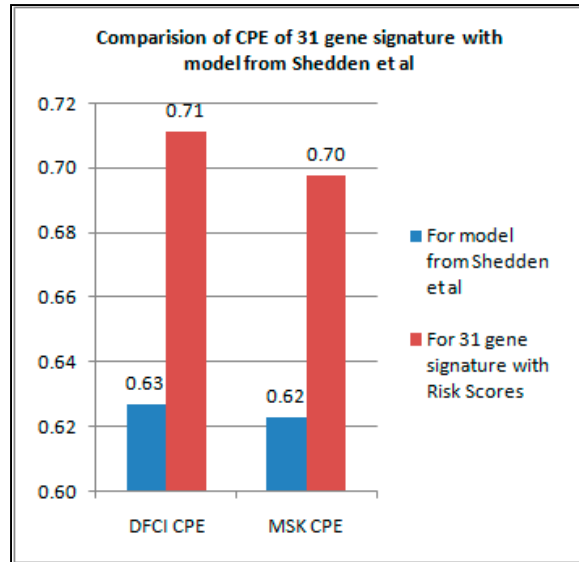


**Figure 4-6: Hazard ratios and 95% Confidence Intervals (obtained from the CPE package which use the risk scores of the entire signature as input) shown along with error bars for 31gene signature and the model from Shedden et al.**



**Figure 4-7: Comparison of p-values (obtained from the CPE package which use the risk scores of the entire signature as input) for 31 gene signature and model from Shedden et al. on a logarithmic scale**

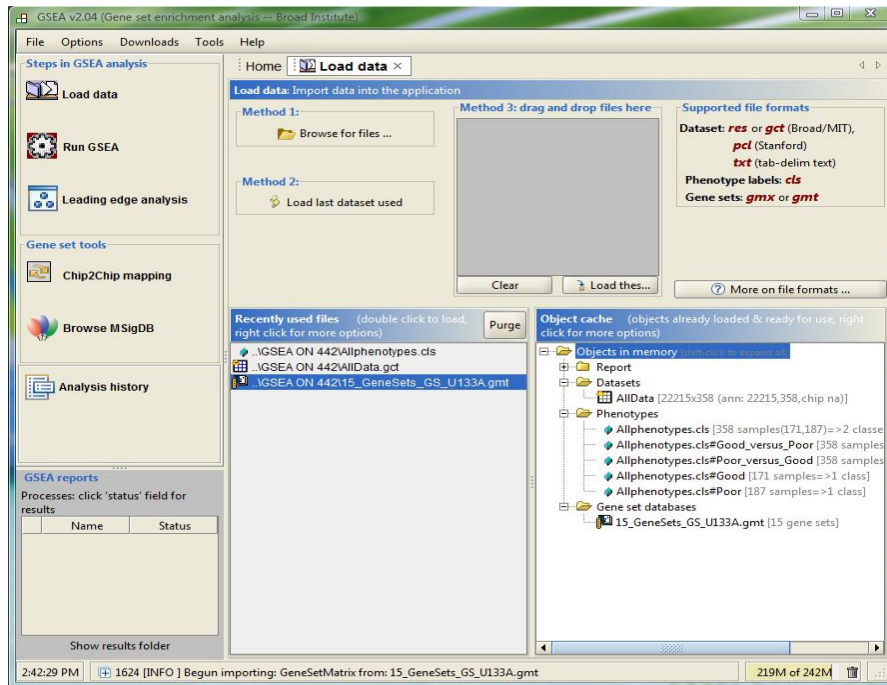




**Figure 4-8 : Concordance Probability Estimates compared between 31 gene signature and the model from Shedden et al.**

### **4.3.3 Gene Set Enrichment Analysis(GSEA)**

To run GSEA [30], gene expression values of all the genes were taken. Since it is better to have as many samples as possible, the training (UM+HLM) and test datasets (DFCI and MSK) were combined to form the 442 samples dataset. The samples were then assigned a class and there were a few censored cases which did not fall in to either of the classes. Hence there were 358 samples after censoring 84 samples. Three files are loaded into GSEA as shown in Figure 4-9: first one is the expression dataset file which contains the gene expression values of the entire set of genes; second one is the phenotype labels file which includes the phenotype labels associated to each sample; third one is the gene sets file which gives the names of genes for one or more gene sets.

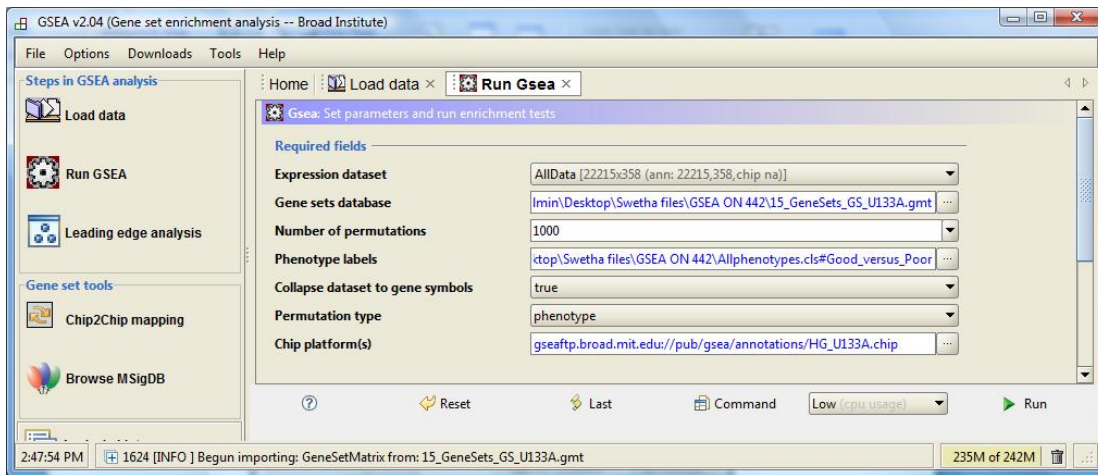


**Figure 4-9: Screenshot for loading data in to GSEA**

We used 15 gene sets which included our 31 gene dataset. The remaining 14 gene sets were extracted from different published papers. The sizes of the datasets used and the sizes that were identified by GSEA are shown in the Table 4-8 below.

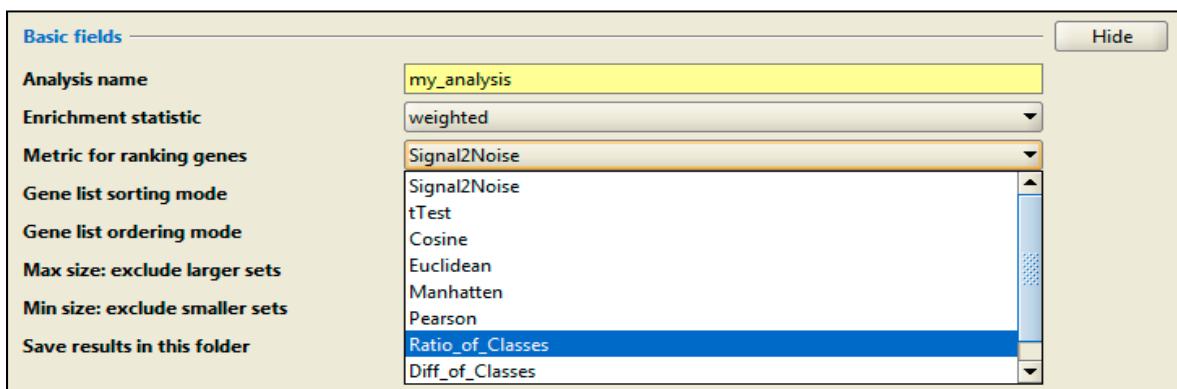
**Table 4-8: Different signatures used to compare the performance of the 31 gene signature in GSEA**

NAME	ORIGINAL SIZE	AFTER RESTRICTING TO DATASET	STATUS
POTTI_133G [51]	131	125	
CHEN_5G [52]	5	5	
BEER_50G [53]	49	44	
SHEDDEN_MA [20]	9591		Rejected!
SHEDDEN_MB [20]	50	38	
SHEDDEN_MC [20]	23	22	
SHEDDEN_MD [20]	36	32	
SHEDDEN_MH [20]	252	223	
BOUTROS_6G [54]	6	6	
BHATTACHARJEE_150G [55]	131	124	
RAPONI_50G [56]	45	39	
LAU_3G [57]	3	3	
LU_64G [58]	63	59	
GUO_35G [59]	35	26	
IMPLICATION_31G	29	25	



**Figure 4-10: Screenshot showing the Basic fields in running GSEA**

The datasets were collapsed to gene symbols and HG\_U133A chip platform was used. 1000 permutations were done on phenotypes as shown in Figure 4-10. Ratio of classes was taken as a metric for ranking the genes and the real values of the genes were considered as shown in Figures 4-11 and 4-12.



**Figure 4-11: Screenshot showing the Selection of Metric for ranking genes**

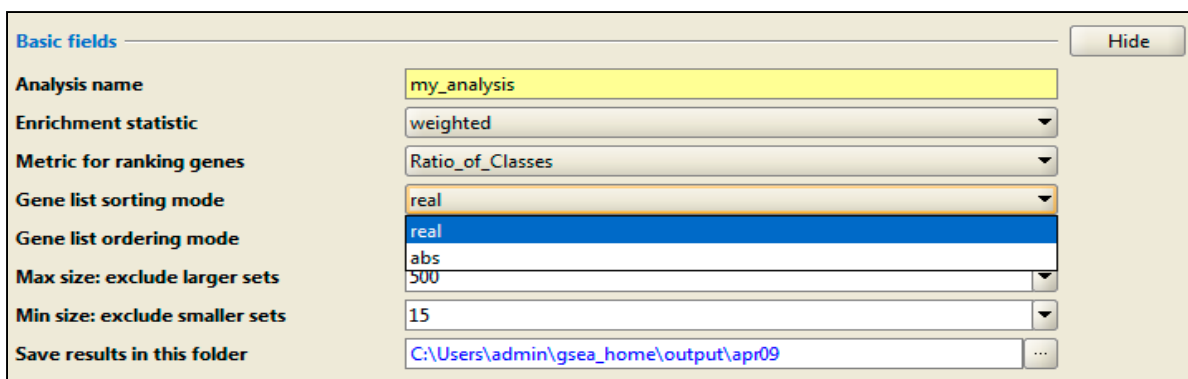


Figure 4-12: Screenshot showing the selection of sorting the gene list based on their real values

There were only 14 gene sets as one of the gene set which had more than 500 genes was filtered out from the analysis. Out of the 14 gene sets, 9 gene sets were enriched in phenotype “Good”. In other words they were upregulated in Good prognosis as shown in Table 4-9. There were 4 gene sets among these 9 gene sets which were significant at FDR<25% which includes the 31genes dataset. Out of the 14 gene sets, there were 5 gene sets which were enriched in the phenotype “Poor”. One of them is significantly enriched at FDR<25% as shown in Table 4-10.

Table 4-9: Different signatures Enriched in phenotype “Good”, which include the 31 gene signature

SIGNATURE INDEX	NAME	SIZE	ES	NES	NOM p-value	FDR q-value	FWER p-value
1	SHEDDEN_MD [20]	32	0.597466	2.385339	0	0	0
2	SHEDDEN_MC[20]	22	0.558553	2.172679	0	0.006652	0.013
3	BHATTACHARJEE_150G [55]	124	0.2782	1.553199	0.044595	0.178669	0.41
4	IMPLICATION_31G	25	0.992265	1.518779	0.255459	0.218715	0.587
5	GUO_35G [59]	26	0.228604	1.359544	0.12989	0.28394	0.845
6	RAPONI_50G [56]	39	0.221071	1.393075	0.156566	0.313299	0.822
7	CHEN_5G [52]	5	0.435005	1.138269	0.299257	0.370527	0.965
8	SHEDDEN_MB [20]	38	0.172793	1.148412	0.299652	0.41296	0.962
9	POTTI_133G [51]	125	0.069968	0.786085	0.694957	0.681874	0.999

Table 4-10: Different signatures Enriched in phenotype “Poor”

SIGNATURE INDEX	NAME	SIZE	ES	NES	NOM p-value	FDR q-value	FWER p-value
10	SHEDDEN_MH [20]	223	-0.51735	-2.33102	0	0.00139	0.002
11	LU_64G [58]	59	-0.14059	-1.22657	0.236994	0.551122	0.664
12	BEER_50G [53]	44	-0.18432	-0.99556	0.416422	0.740632	0.897
13	BOUTROS_6G [54]	6	-0.2156	-0.72	0.826552	0.799505	0.989
14	LAU_3G [57]	3	-0.36642	-0.81345	0.699589	0.848576	0.977

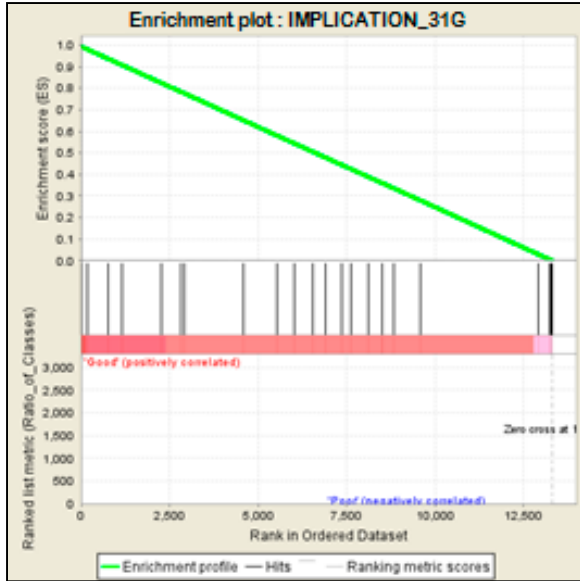


Figure 4-13: Enrichment score plot for the 31 gene signature picked from implication networks which shows the Enrichment profile on the top and the ranked list metric on the bottom

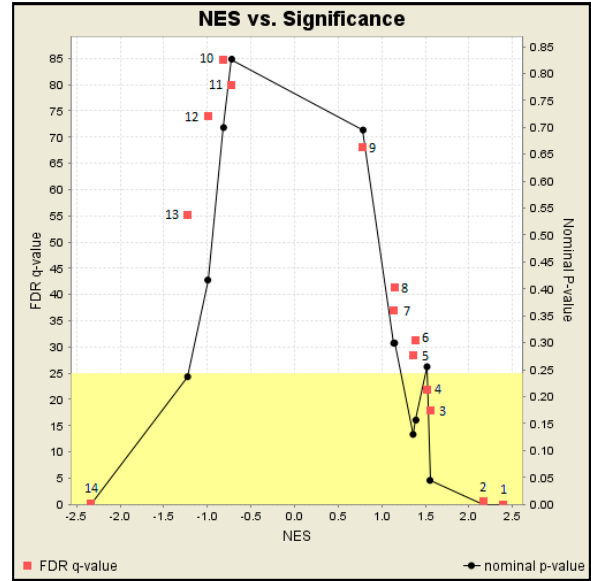


Figure 4-14: Plot showing the Nominal Enrichment Scores, False Discovery Rates and Nominal P-values for all the signatures with Signature index of each signature from Table 4-9 and 4-10. Index 4 represents the 31-gene signature from implication networks

The Figures 4-13 above shows the Enrichment score plot for the 31 gene signature and Figure 4-14 shows the comparison plot for nominal p-values and FDR with respect to NES for all the signatures used, highlighting the 31 gene signature.

#### 4.3.4 Comparison of model with other classification methods using WEKA

The model used to classify the samples in to two groups was compared against randomly picked classifiers in Weka. Five classifiers were considered which are Random Tree Classifier, Support Vector Machine (SVM or SMO) classifier, K-nearest neighbors (IBK) classifier, Multilayered Perceptron classifier (Neural networks), and Bayes Net classifier.

The actual classification of the groups based on the survival period and status of post operative survival was compared with the classifications performed by the above mentioned classifiers. The classification accuracies using nearest point classification in Cox model for implication networks were calculated from the sensitivity, specificity, and overall accuracy measures mentioned in the previous sections. Bayesian networks generated from TETRAD IV were also compared for overall accuracies.

The comparison was performed for 5 year survival on the Training dataset (ULM/HLM). This dataset had 229 samples after 27 samples were censored. The results are given in the Table 4-11 below. It can be seen that the Cox model had the best classification accuracy. The sensitivity and specificity values were also calculated along with the overall accuracy. Significance test was conducted on the accuracies obtained to get the p-values. The significance of all the other models was calculated with accuracy of the Cox model as reference. If the p-values are small they imply that the NULL hypothesis (here equal significance of the two compared overall accuracies) is rejected and that the Alternative hypothesis (significance of larger overall accuracy > significance of smaller overall accuracy) is strongly supported. Since all the p-values are very small, it is obvious that the Cox model is highly significant when compared to other models. These classifications were used to find the Concordance Probability Estimates, log-rank p-values, hazard ratios, and confidence intervals for each model.

**Table 4-11: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the training dataset using p-values from significance test**

Training dataset	correctly classified instances	incorrectly classified instances	Sensitivity	Specificity	Overall Accuracy	Z-score	P-value
Random Tree	121	108	56.8%	48.1%	52.8%	2.48	0.006
SMO (or SVM)	132	97	72.0%	40.4%	57.6%	1.45	0.073
IBK	129	100	58.4%	53.8%	56.3%	1.73	0.042
Multilayered Perceptron(Neural)	132	97	64.0%	50.0%	57.6%	1.45	0.073
Bayes Net	120	109	90.4%	6.7%	52.4%	2.56	0.005
Bayesian Networks using TETRAD IV	133	96	65.6%	49.0%	58.1%	1.34	0.090
Implication Networks - COX model	147	82	66.4%	61.5%	64.2%	-	-

The Concordance Probability Estimates, log-rank p-values, hazard ratios, and the confidence interval calculated from classifications from various models on training dataset (ULM/HLM) are shown below in Table 4-12. The CPE values are supposed to be higher than 0.5 and the higher they are, the better is the model. The log-rank p-values of the models must be less than 0.05 and the lesser they are, the more significant the model is considered to be. The hazard ratios should be as high as possible and the 95% confidence intervals should not contain 1 in their range. The Cox model had highly significant results when compared to the other techniques.

**Table 4-12: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the training dataset**

Training dataset	CPE	log-rank p-values	hazard ratios with 95% CI
Random Tree	0.524456	0.535865	1.1 [0.809, 1.5]
SMO (or SVM)	0.539067	0.157513	1.27 [0.909, 1.77]
IBK	0.55727	0.145277	1.26 [0.923, 1.72]

Multilayered Perceptron(Neural)	0.57637	0.053659	1.36 [0.992, 1.87]
Bayes Net	0.516962	0.816327	0.934 [0.53, 1.65]
Implication Networks-COX model	0.7123937	3.90E-08	2.48 [1.79, 3.42]

Since the training dataset had good results, we tried to validate the model on the testing datasets. The classification from survival time and status was used as actual classification for the DFCI test dataset. This dataset had 64 samples after 18 samples were censored. The classification results for DFCI dataset for various classifiers used in Weka are shown below in comparison with the Cox model in Table 4-13. Bayesian networks generated from TETRAD IV were also compared for overall accuracies. The sensitivity and specificity values were also calculated along with the overall accuracy. Significance test was conducted on the accuracies obtained to get the p-values. The significance of all the other models was calculated with accuracy of the Cox model as reference. If the p-values are small they imply that the NULL hypothesis (here equal significance of the two compared overall accuracies) is rejected and that the Alternative hypothesis (significance of larger overall accuracy > significance of smaller overall accuracy) is strongly supported. Since all the p-values are very small, it is obvious that the Cox model is highly significant when compared to other models.

**Table 4-13: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the DFCI test dataset using p-values from significance test**

DFCI dataset	correctly classified instances	incorrectly classified instances	Sensitivity	Specificity	Overall Accuracy	Z-score	P-value
Random Tree	34	30	64.3%	44.4%	53.1%	1.44	0.075
SMO (or SVM)	27	37	92.9%	2.8%	42.2%	2.66	0.004
IBK	26	38	39.3%	41.7%	40.6%	2.83	0.002
Multilayered Perceptron (Neural)	27	37	92.9%	2.8%	42.2%	2.66	0.004
Bayes Net	28	36	100.0%	0.0%	43.8%	2.48	0.007



Bayesian Networks using TETRAD IV	30	34	60.7%	36.1%	46.9%	2.13	0.017
Implication Networks-Cox Model	42	22	92.9%	44.4%	65.6%	-	-

The Concordance Probability Estimates, log-rank p-values, hazard ratios, and the confidence interval calculated from the model on DFCI dataset are shown below. All the instances in Bayes Net were classified to the same group (Poor prognosis) and hence calculation of the parameters was not possible (NA). This is because the calculation of parameters requires at least two different groups. The CPE values are supposed to be higher than 0.5 and the higher they are, the better is the model. The log-rank p-values of the models must be less than 0.05 and the lesser they are, the more significant the model is considered to be. The hazard ratios should be as high as possible and the 95% confidence intervals should not contain 1 in their range. From the results shown below in Table 4-14, it can be concluded that Cox model had much significant results when compared to other techniques.

**Table 4-14: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the DFCI test dataset**

DFCI dataset	CPE	log-rank p-values	hazard ratios with 95% CI
Random Tree	0.543057	0.620414	1.19 [0.597, 2.36]
SMO (or SVM)	0.693209	0.319597	0.443 [0.106, 1.85]
IBK	0.600628	0.230398	0.665 [0.34, 1.3]
Multilayered Perceptron(Neural)	0.693209	0.319597	0.443 [0.106, 1.85]
Bayes Net	NA	NA	NA
Implication Networks-COX model	0.845703	0.0014	5.48 [1.93, 15.6]

The classification from survival time and status was used as actual classification for the MSK test dataset. This dataset had 65 samples after 39 samples were censored. The classification results for MSK dataset for various classifiers used in Weka are shown below in Table 4-15. Bayesian networks generated from TETRAD IV were also compared for overall accuracies. The sensitivity and specificity values were also calculated along with the overall accuracy. Significance test was conducted on the accuracies obtained to get the p-values. The significance of all the other models was calculated with accuracy of the Cox model as reference. If the p-values are small they imply that the NULL hypothesis (here equal significance of the two compared overall accuracies) is rejected and that the Alternative hypothesis (significance of larger overall accuracy > significance of smaller overall accuracy) is strongly supported. Since all the p-values are very small, it is obvious that the Cox model is highly significant when compared to other models.

**Table 4-15: Comparison of classification accuracies of various methods from Weka with Cox model on implication networks on the MSK test dataset using p-values from significance test**

MSK dataset	correctly classified instances	incorrectly classified instances	Sensitivity	Specificity	Overall Accuracy	Z-score	P-value
Random Tree	35	30	67.6%	38.7%	53.8%	1.44	0.075
SMO (or SVM)	31	34	2.9%	96.8%	47.7%	2.13	0.017
IBK	26	39	29.4%	51.6%	40.0%	2.99	0.001
Multilayered Perceptron(Neural)	34	31	100.0%	0.0%	52.3%	1.61	0.054
Bayes Net	34	31	100.0%	0.0%	52.3%	1.61	0.054
Bayesian Networks using TETRAD IV	33	32	58.8%	41.9%	50.8%	1.78	0.037
Implication Networks-COX model	43	22	64.7%	67.7%	66.2%	-	-

The log-rank p-values, CPE, hazard ratios, and the confidence interval calculated from the model on MSK dataset are shown below. All the instances in Multilayered Perceptron and Bayes Net were classified to the same group (Poor prognosis) and hence calculation of the parameters was not possible (NA). This is because the calculation of parameters requires at least two different groups. The CPE values are supposed to be higher than 0.5 and the higher they are, the better is the model. The log-rank p-values of the models must be less than 0.05 and the lesser they are, the more significant the model is considered to be. The hazard ratios should be as high as possible and the 95% confidence intervals should not contain 1 in their range. From the results shown below in Table 4-16, it can be concluded that Cox model had much significant results when compared to other techniques.

**Table 4-16: Comparison of Concordance Probability Estimates, log-rank p-values (obtained from the CPE package with risk scores of the entire signature as input), hazard ratios (based on 5-year cutoff) and confidence intervals (obtained from the CPE package with risk scores of the entire signature as input) of various methods from Weka with Cox model on implication networks on the MSK test dataset**

MSK dataset	CPE	log-rank p-values	hazard ratios with 95% CI
Random Tree	0.539067	0.651084	1.17 [0.59, 2.32]
SMO (or SVM)	0.538504	0.882016	1.17 [0.16, 8.53]
IBK	0.548117	0.565333	0.824 [0.425, 1.6]
Multilayered Perceptron(Neural)	NA	NA	NA
Bayes Net	NA	NA	NA
Implication Networks-COX model	0.782544	0.00019	3.6 [1.84, 7.05]

Thus all the results above show that the Cox model predicted the outcomes with best classification accuracies, CPE, log-rank p-values, and hazard ratios with 95% confidence intervals.

#### 4.4 Clinical Evaluation

Clinical evaluation was done by comparing various predictive factors such as Age, Gender, Lymph node metastasis, Tumor size, etc with the risk scores from our model. This is done using multivariate Cox proportional hazards model.

##### Multivariate Analysis on Cox proportional hazards model

Multivariate analysis was done using the Cox proportional hazards model to compare the significance of the risk scores from 31 gene signature with the other pathological factors. For multivariate analysis, the risk scores of the 31 gene signature obtained in Cox model earlier were used as a predictor. Mostly used covariates such as Age, Gender, Lymph node Metastasis, and Tumor size were used with the risk scores of 31 gene predictors as shown in Table 4-17. In this table, the other predictors were fit in to Cox model without the risk scores and then they were again fit in to Cox model with the risk scores included. Both the analyses were compared which showed that the addition of risk scores to other predictors made the significance of other predictors to decrease and that the risk scores of our model had the most significant p-value.

**Table 4-17: Multivariate Cox Proportional Analysis of Age, Gender, Lymph node Metastasis, Tumor size and Risk Score\***

Variable	Log-rank p-value	Hazard Ratio [95% CI] <sup>‡</sup>
<b>Analysis without risk score</b>		
AGE	0.00081	1.69 [1.243, 2.3]
GENDER	0.059	0.777 [0.598, 1.01]
Lymph node Metastasis	6.20E-14	2.716 [2.092, 3.53]
Tumor Size	0.0035	1.537 [1.151, 2.05]
<b>Analysis with risk score</b>		
31 genes Risk Scores	2.30E-14	2.43 [1.933, 3.05]
AGE	0.0056	1.55 [1.136, 2.11]
GENDER	0.12	0.81 [0.623, 1.05]
Lymph node Metastasis	1.00E-13	2.7 [2.081, 3.51]
Tumor Size	0.084	1.29 [0.966, 1.73]

**\*Age was a binary variable (0 for an age less than 60 years and 1 for an age of 60 years or greater); Gender was a binary variable (0 for Male and 1 for Female); Lymph node Metastasis was a binary variable (0 for N0-stage and 1 for all other N-stages and missing values); Tumor size was a binary variable (0 for T0-stage and 1 for all other T-stages and missing values); Risk score was a continuous variable.**  
**‡CI denotes Confidence interval.**

To perform a complete and comprehensive analysis on the pathological factors of Non-Small Cell Lung Cancer, all the other covariates like Race, Smoking status, and Tumor grade were added to the above covariates to find the log-rank p-values and Hazard ratios as shown in Table 4-18. Again analyses with and without risk scores was done and it can be seen that 31 gene predictors are most significant. The smoking status can be ignored as it has a group of unknown samples which would not allow the correct prediction of the entire group. The descriptions of each of the variables are given in the legends of the tables.

**Table 4-18: Multivariate Cox Proportional Analysis of Age, Gender, Race, Smoking Status, Lymph node Metastasis, Tumor size, Tumor grade and Risk Score\***

Variable	Log-rank p-value	Hazard Ratio [95% CI] <sup>‡</sup>
<b>Analysis without risk score</b>		
AGE	0.00069	1.705 [1.253, 2.32]
GENDER	0.059	0.763 [0.576, 1.01]
RACE		
Other	0.76	0.877 [0.375, 2.05]
White	0.72	1.161 [0.512, 2.63]
SMOKING STATUS		
Smokers	0.4	1.23 [0.761, 1.99]
Unknown	0.25	1.385 [0.797, 2.41]
Lymph node Metastasis	3.60E-14	2.788 [2.138, 3.64]
Tumor Size	0.0026	1.569 [1.17, 2.1]
TUMOR GRADE		
POORLY DIFFERENTIATED	0.35	1.144 [0.865, 1.51]
WELL DIFFERENTIATED	0.38	1.211 [0.788, 1.86]
<b>Analysis with risk score</b>		
31 genes Risk Scores	1.80E-13	2.403 [1.903, 3.03]
AGE	0.0061	1.544 [1.132, 2.11]
GENDER	0.19	0.827 [0.621, 1.1]
RACE		
Other	0.56	0.774 [0.329, 1.82]
White	0.74	0.872 [0.382, 1.99]
SMOKING STATUS		
Smokers	0.54	1.164 [0.719, 1.88]
Unknown	0.3	1.35 [0.769, 2.37]
Lymph node Metastasis	8.90E-14	2.737 [2.1, 3.57]
Tumor Size	0.074	1.311 [0.975, 1.76]
TUMOR GRADE		
POORLY DIFFERENTIATED	0.44	1.117 [0.843, 1.48]
WELL DIFFERENTIATED	0.61	1.116 [0.727, 1.71]

\*Age was a binary variable (0 for an age less than 60 years and 1 for an age of 60 years or greater); Gender was a binary variable (0 for Male and 1 for Female); Race was a binary variable (Other relative to Black/African American, White relative to Black/African American; Other includes a few Native Hawaiian/Pacific Islander, Asian and unknown); Smoking status was a binary variable (Smokers relative to Non-Smokers and Unknown status relative to Non-Smokers); Lymph node Metastasis was a binary variable (0 for N0-stage and 1 for all other N-stages and missing values); Tumor size was a binary variable (0 for T0-stage and 1 for all other T-stages and missing values); Tumor grade was a binary variable (Poorly differentiated relative to moderately differentiated and Well differentiated relative to moderately differentiated); Risk score was a continuous variable.

<sup>‡</sup> CI denotes Confidence interval.

## 4.5 Topological Validation

To derive the biological insight using curated databases, topological validation was performed on the prognostic signature obtained from prediction logic based implication networks. This validation also required the comparison of the biological relevance of the interactions present in the implication network with a currently used network such as Bayesian network. There are many techniques for structural validation of the gene signature.

In prognostic validation, the best probe was considered based on the minimum p-value after fitting in to Cox model. But in structural validation, the average of all the duplicate probes was taken and was used for the analysis. The different structural validation techniques used include Prodistin, Kegg, NCI pathways, PubMed interactions, Matisse, String 8, Ingenuity Pathway Analysis, and Pathway Studio. Tetrad IV was used to generate Bayesian networks which were compared in different aspects with the implication network.

The implication network was built from the 31 genes and the hallmarks used to identify the signature.

The gene expression data of the 22215 genes was sorted according to their gene symbols. The averages of the duplicate probes were taken which leaves 13658 unique genes. The 31 genes along with the hallmarks which were used to get the 31 gene signature were picked. Hence there were 31 genes plus 8 hallmarks. There were 256 samples in the Training dataset. This data is split in to 2 files, Metastasis (high risk) and Non-Metastasis (low risk) groups, based on the number of months they survived and survival status. If the number of months the patients survived was greater than 60 (5 years), the sample was put in Non-Metastasis group (low risk). If the number of months the patients survived was less than 60 months and if it was known that the

patient died, the sample was put in Metastasis group (high risk). If the number of months the patients survived was less than 60 months and if it was not known whether the patient died, the sample was censored.

The Metastasis group (high risk) had 125 samples and the Non-Metastasis group (low risk) had 104 samples. The remaining 27 samples were censored as shown in Table 4-19. The data in the files was converted in to 1's and 0's by partitioning based on the mean which was used to generate the interactions among genes.

**Table 4-19: Number of patients in each of the groups in each dataset along with number of censored patients**

	# patients in high risk (Metastasis) group	# patients in low risk (Non-Metastasis) group	# patients censored
Training dataset	125	104	27
DFCI dataset	28	36	18
MSK dataset	34	31	39

Interactions between genes were generated using the files which had binary data. There were 1021 interactions from Metastasis group (high risk) and 897 interactions from the Non-Metastasis group (low risk) as shown in Table 4-20.

The above steps were repeated for the 31 gene signature in DFCI data set and the MSK data set.

The DFCI dataset had 82 samples. After partitioning, there were 28 samples for Metastasis group (high risk) and 36 samples for Non-Metastasis group (low risk). There were 18 samples which were censored as shown in Table 4-19. There were 787 interactions from Metastasis group (high risk) and 938 interactions from Non-Metastasis group (low risk) as shown in Table 4-20.



The MSK dataset had 104 samples. After partitioning, there were 34 samples for Metastasis group (high risk) and 31 samples for Non-Metastasis group (low risk). There were 39 samples that were censored as shown in Table 4-19. There were 992 interactions from Metastasis group (high risk) and 996 interactions from Non-Metastasis group (low risk) as shown in Table 4-20.

**Table 4-20: Number of interactions between the 31 genes and the 8 hallmarks for various datasets in both the groups**

Low risk (Non-Metastasis)	High risk (Metastasis)
Interactions from Training=897	Interactions from Training=1021
Interactions from DFCI=938	Interactions from DFCI=787
Interactions from MSK=996	Interactions from MSK=992

Differential components are the interactions that are present in one group (high or low) but not present in the other group (low or high).

The interactions from the good and poor prognosis of the Training dataset had 235 interactions in common. So there were 786 interactions from the Metastasis group (high risk) and 662 interactions from the Non-Metastasis group (low risk) that were considered as the differential components as shown in Table 4-21.

Similarly, the interactions from the good and poor prognosis of the DFCI dataset had 308 interactions in common. So there were 479 interactions from the Metastasis group (high risk) and 630 interactions from the Non-Metastasis group (low risk) that were considered as the differential components as shown in Table 4-21.

Similarly, the interactions from the good and poor prognosis of the MSK dataset had 359 interactions in common. So there were 633 interactions from the Metastasis group (high risk) and

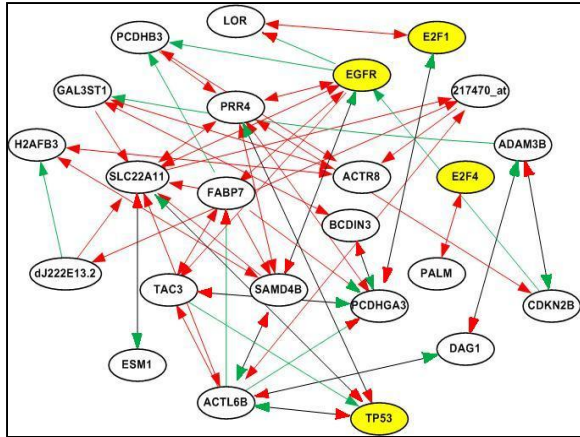
637 interactions from the Non-Metastasis group (low risk) that were considered as the differential components as shown in Table 4-21.

**Table 4-21: Number of differential components between both the groups for the 31 genes and the 8 hallmarks for various datasets**

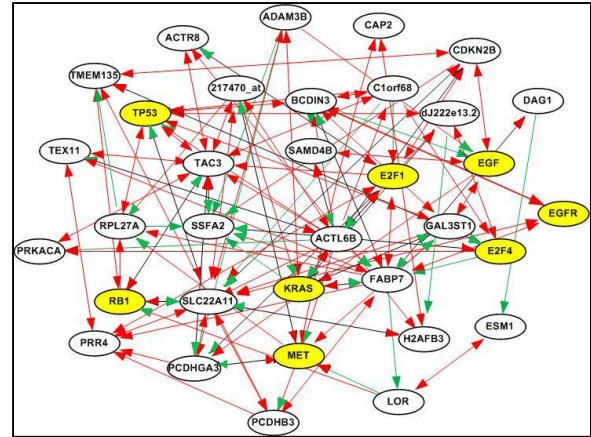
Low risk (Non-Metastasis)	High risk (Metastasis)
Differential Components from Training=662	Differential Components from Training=786
Differential Components from DFCI=630	Differential Components from DFCI=479
Differential Components from MSK=637	Differential Components from MSK=633

After getting the differential components for each dataset, the interactions that were common among every two datasets and also those interactions that were common among all the three datasets were found.

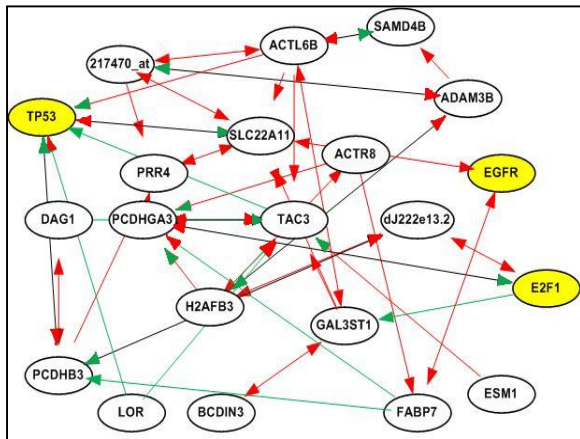
Among the Metastasis group (high risk), there were 81 interactions common to the differential components of the Training dataset and the DFCI dataset. The interactions between the genes are shown graphically in the Figure 4-15 below. There were 168 interactions common to the differential components of the Training dataset and the MSK dataset. The interactions between the genes are shown graphically in the Figure 4-16 below. There were 61 interactions common to the differential components of the DFCI dataset and the MSK dataset. The interactions between the genes are shown graphically in the Figure 4-17 below. The genes in yellow color are the Hallmarks used and the uncolored genes are the regular signature genes. There were 31 interactions that were common to all the three datasets in the Metastasis group (high risk). The interactions between the genes are shown in the Figure 4-18 below.



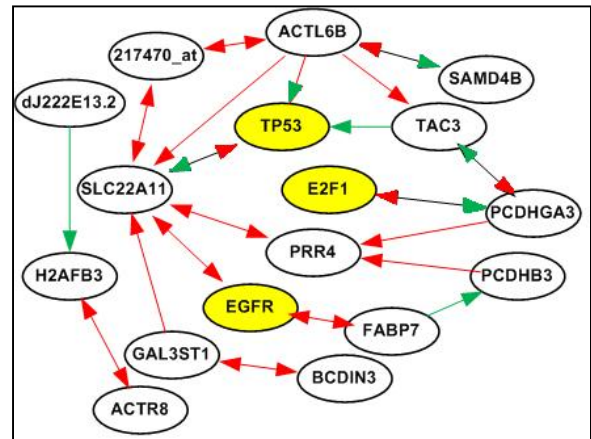
**Figure 4-15: Differential Components common to Train & DFCI datasets in high risk group**



**Figure 4-16: Differential Components common to Train & MSK datasets in high risk group**



**Figure 4-17: Differential Components common to DFCI & MSK datasets in high risk group**

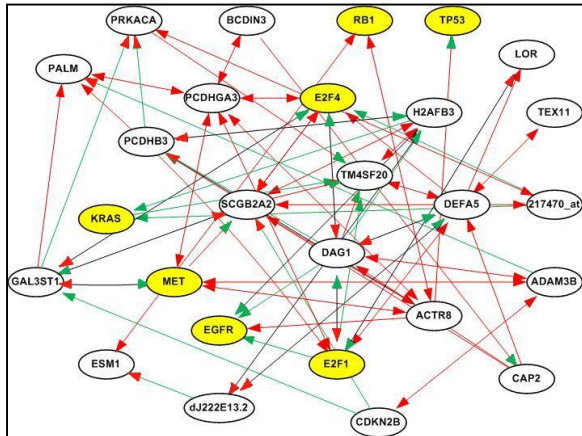


**Figure 4-18: Differential Components common to all 3 datasets in high risk group**

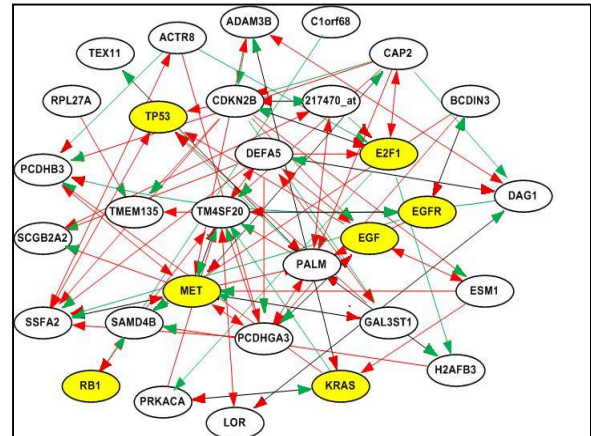
Among the Non-Metastasis group (low risk), there were 96 interactions common to the differential components of the Training dataset and the DFCI dataset. The interactions between the genes are shown graphically in the Figure 4-19 below. There were 106 interactions common to the differential components of the Training dataset and the MSK dataset. The interactions between the genes are shown graphically in the Figure 4-20 below. There were 82 interactions common to the differential components of the DFCI dataset and the MSK dataset. The

interactions between the genes are shown graphically in the Figure 4-21 below. There were 27 interactions that were common to all the three datasets in the Non-Metastasis group (low risk).

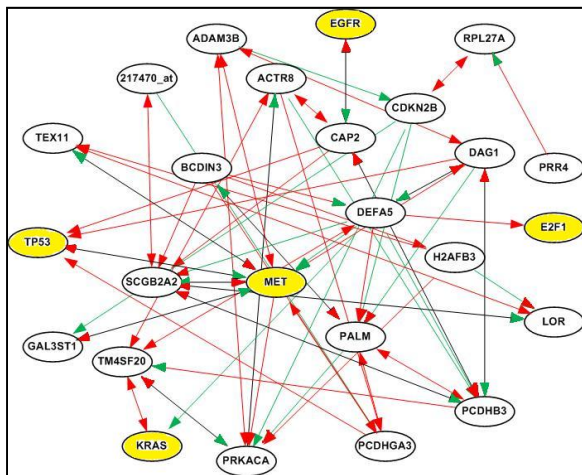
The interactions between the genes are shown in the Figure 4-22 below.



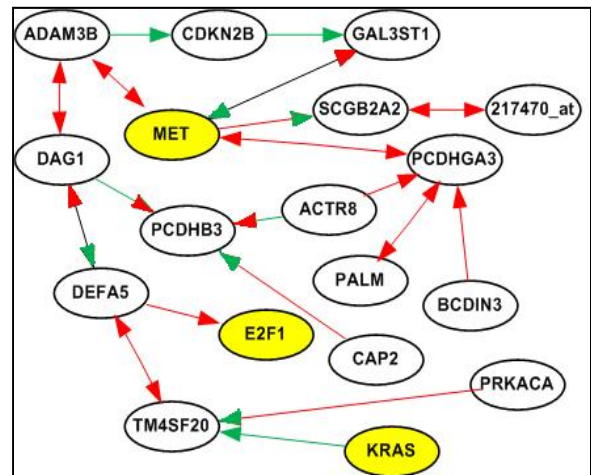
**Figure 4-19: Differential Components common to Train & DFCI datasets in low risk group**



**Figure 4-20: Differential Components common to Train & MSK datasets in low risk group**



**Figure 4-21: Differential Components common to DFCI & MSK datasets in low risk group**

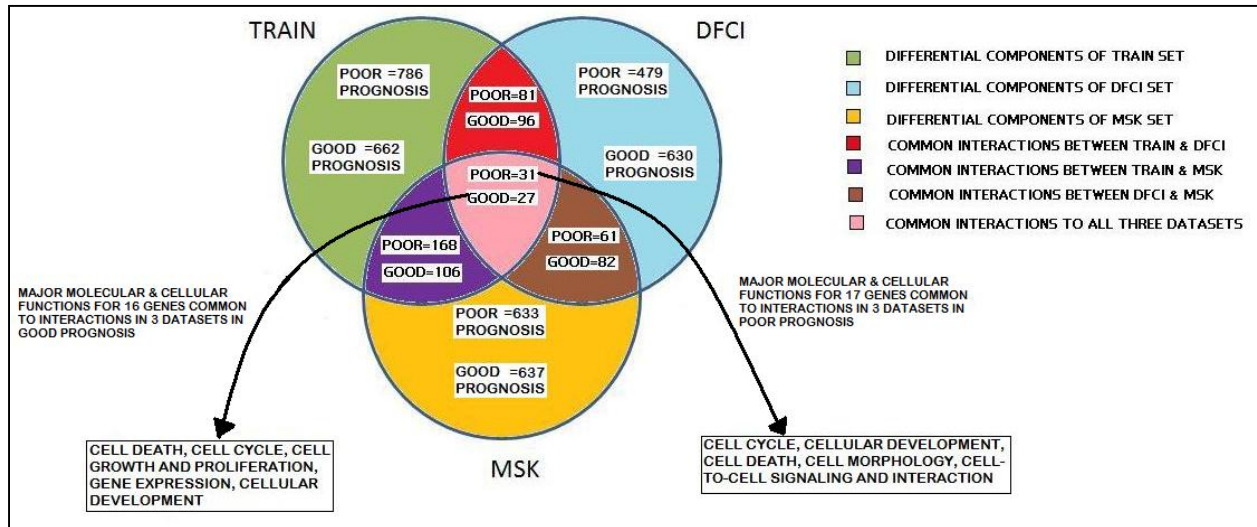


**Figure 4-22: Differential Components common to all three datasets in low risk group**

All these common interactions between the three datasets are shown in the Figure 4-23 below.

This figure also shows the number of genes present in the interactions common to all the three

datasets in both poor and good prognosis groups. The molecular and cellular functions of these genes are mentioned which were extracted from Ingenuity Pathway Analysis (IPA)



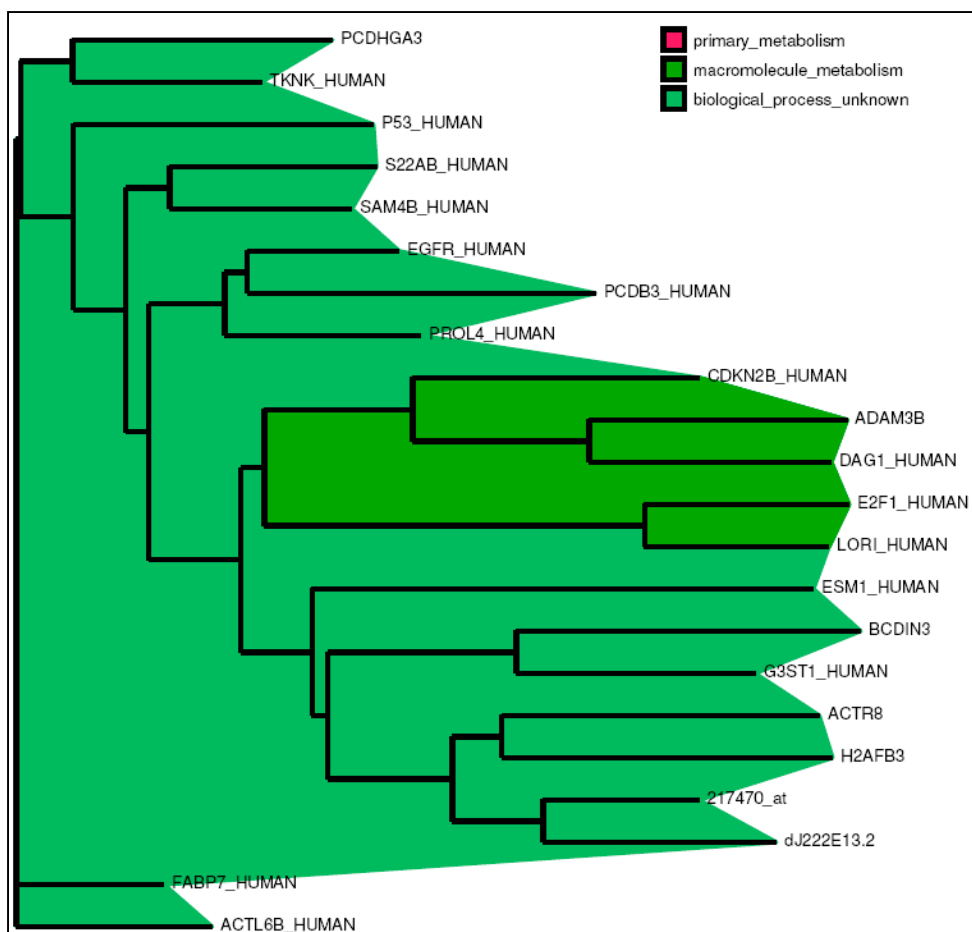
**Figure 4-23: Differential Components common among the three datasets in both the prognosis groups where good prognosis corresponds to low risk group and poor prognosis corresponds to high risk group and the major molecular and cellular functions identified from IPA were also shown**

#### 4.5.1 PRODISTIN

PRODISTIN is web based software that functionally classifies the genes based on the protein-protein interactions. It is based on the principle that the more two proteins share common interactors, the more they are functionally related. It clusters proteins in to functional classes depending whether they participate in the same cellular process or not. It also predicts function for unknown genes.

The process started with the selection of species to Homo sapiens. Then the interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to Training data and the DFCI data in the Metastasis group were loaded. The

gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 22 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 22 classified genes, there were 12 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 3 different GO terms which are shown in the Figure 4-24 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



**Figure 4-24: Clustering from interactions Common to Train and DFCI Metastasis group from PRODISTIN**

The terms primary metabolism and macromolecule metabolism fall in to one class which is give a class number 1. P-values are shown in Table 4-22. The p-values are not very significant in this dataset.

**Table 4-22: p-values of Gene Ontology terms identified from known classes in Common interactions among Train and DFCI Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	5	primary metabolism	0.4762
1	5	macromolecule metabolism	0.4762

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to Training data and the MSK data in the Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 34 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 34 classified genes, there were 16 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 10 different GO terms which are shown in the Figure 4-25 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.





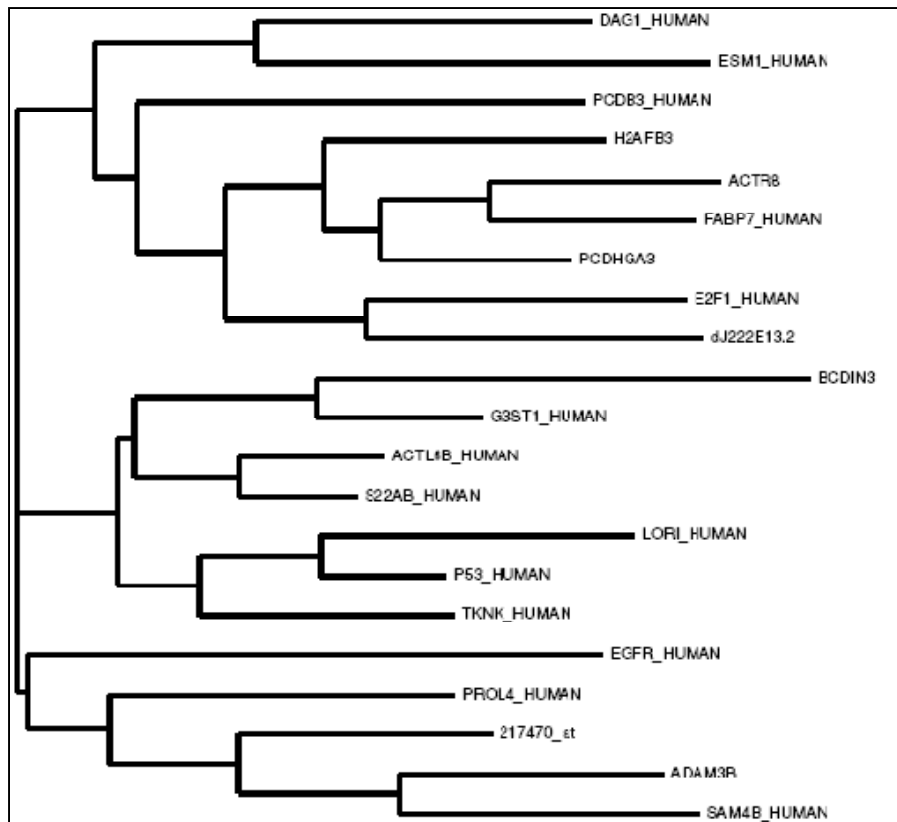
**Figure 4-25: Clustering from interactions Common to Train and MSK Metastasis group from PRODISTIN**

The terms cell communication, protein modification and system development fall into class number 1, primary metabolism and macromolecule metabolism fall in to class number 2, protein metabolism and cellular metabolism fall in to class number 3, signal transduction falls in to class number 4, and morphogenesis falls in to class number 7. P-values are shown in Table 4-23. The p-values of primary metabolism and macromolecule metabolism are significant.

**Table 4-23: p-values of Gene Ontology terms identified from known classes in Common interactions among Train and MSK Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	5	cell communication	0.3921
1	5	protein modification	0.2247
1	5	system development	0.2941
2	12	primary metabolism	0.0498
2	12	macromolecule metabolism	0.0498
3	8	protein metabolism	0.3628
3	8	cellular metabolism	0.2639
4	6	signal transduction	0.3620
7	6	morphogenesis	0.3111

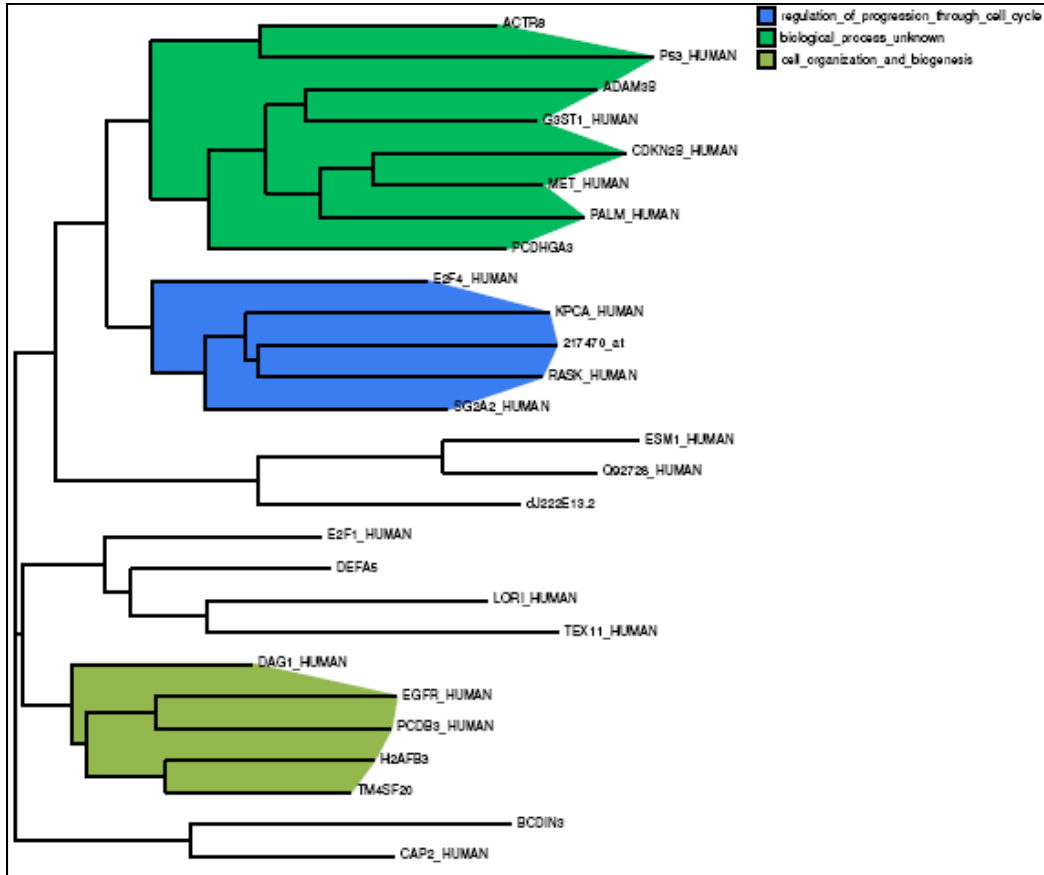
The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to DFCI data and the MSK data in the Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 21 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 21 classified genes, there were 11 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were no GO terms identified. This is shown in Figure 4-26 below. Hence there are no p-values identified. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



**Figure 4-26: Clustering from interactions Common to DFCI and MSK Metastasis group from PRODISTIN**

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to Training data and the DFCI data in the Non-Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 27 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 27 classified genes, there were 11 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 3 different GO terms which are shown in the Figure 4-27 below. If there are

multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



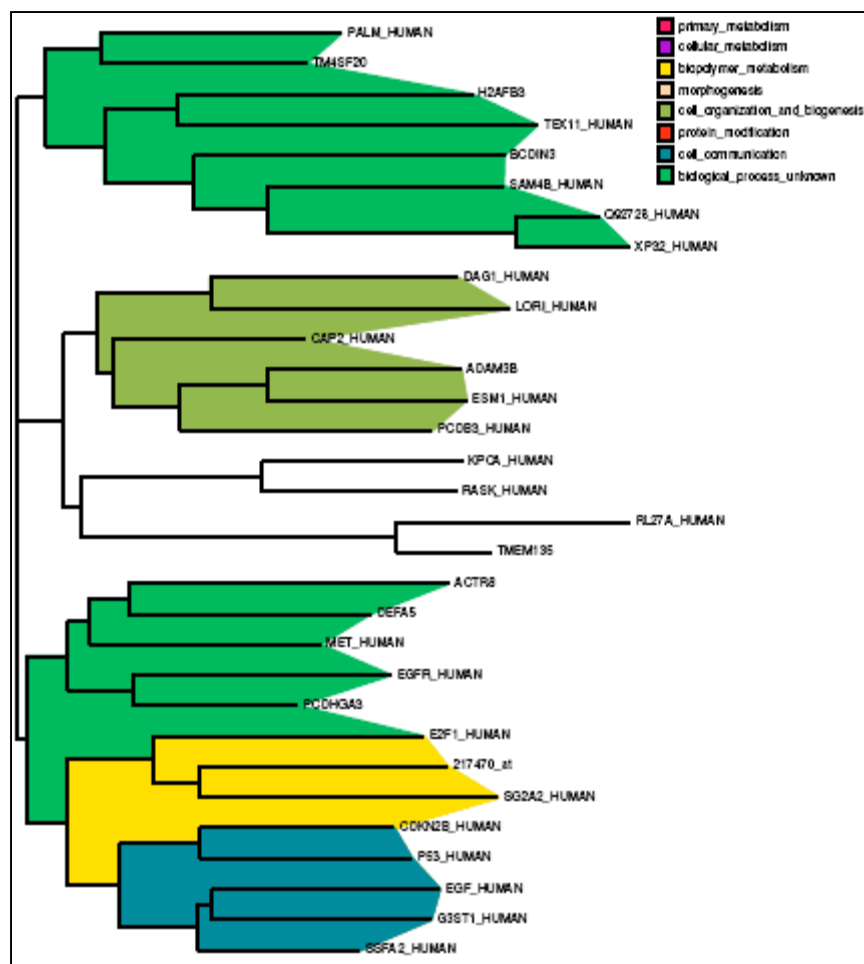
**Figure 4-27: Clustering from interactions Common to Train &DFCI Non-Metastasis group from PRODISTIN**

The term regulation of progression through cell cycle falls in to class number 1 and cell organization and biogenesis falls in to class number 3. P-values are shown in Table 4-24. The p-values are not very significant in this dataset.

**Table 4-24: p-values of Gene Ontology terms identified from known classes in Common interactions among Train &DFCI Non-Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	5	regulation of progression through cell cycle	0.2398
3	5	cell organization and biogenesis	0.3916

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to Training data and the MSK data in the Non-Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 31 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 31 classified genes, there were 14 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 8 different GO terms which are shown in the Figure 4-28 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



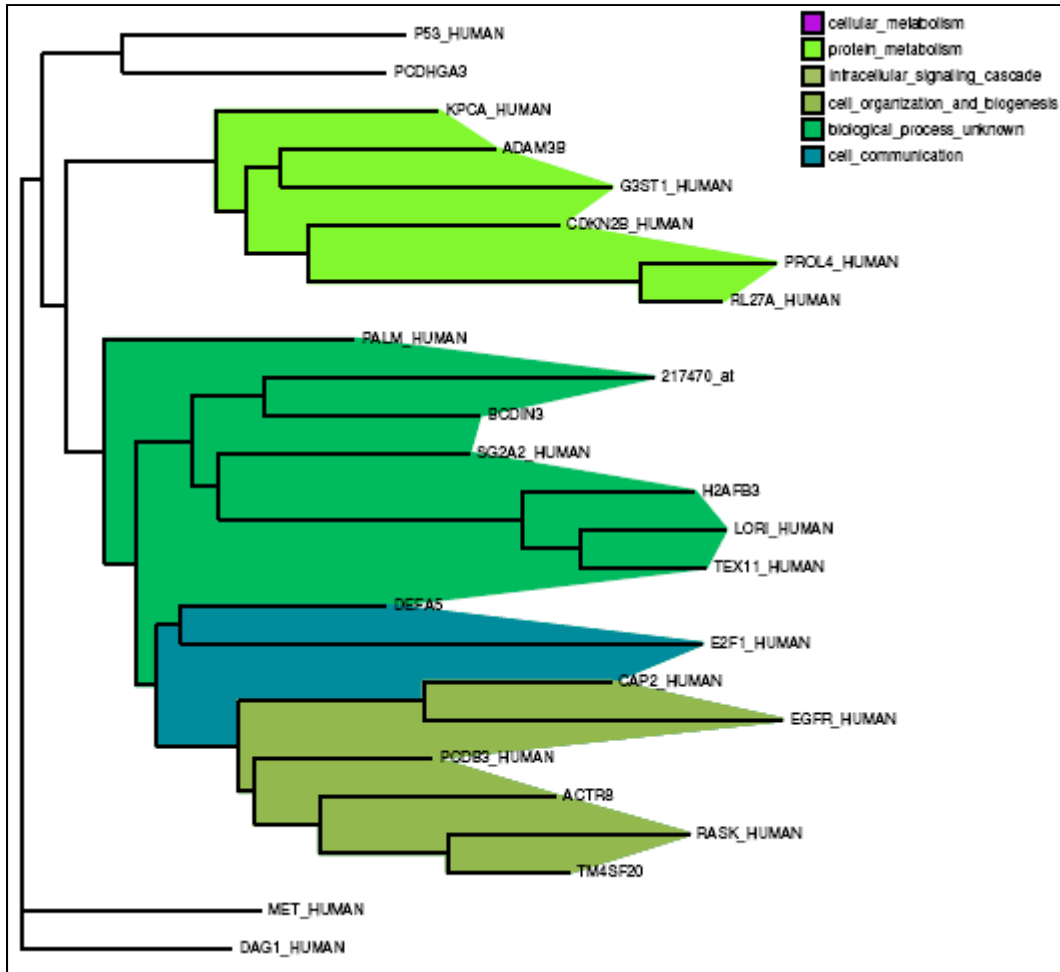
**Figure 4-28: Clustering from interactions Common to Train & MSK Non-Metastasis group from PRODISTIN**

The terms primary metabolism, cellular metabolism and biopolymer metabolism fall into class number 1, morphogenesis, cell organization, and biogenesis fall into class number 2 and protein modification and cell communication fall into class number 3. P-values are shown in Table 4-25. The p-values are very significant in this dataset.

**Table 4-25: p-values of Gene Ontology terms identified from known classes in Common interactions among Train & MSK Non-Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	8	primary metabolism	0.1282
1	8	cellular metabolism	0.2447
1	8	biopolymer metabolism	0.3426
2	6	morphogenesis	0.1573
2	6	cell organization and biogenesis	0.2447
3	5	protein modification	0.2884
3	5	cell communication	0.4038

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to DFCI data and the MSK data in the Non-Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 25 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 25 classified genes, there were 10 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 6 GO terms as shown in the Figure 4-29 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



**Figure 4-29: Clustering from interactions Common to DFCI & MSK Non-Metastasis group from PRODISTIN**

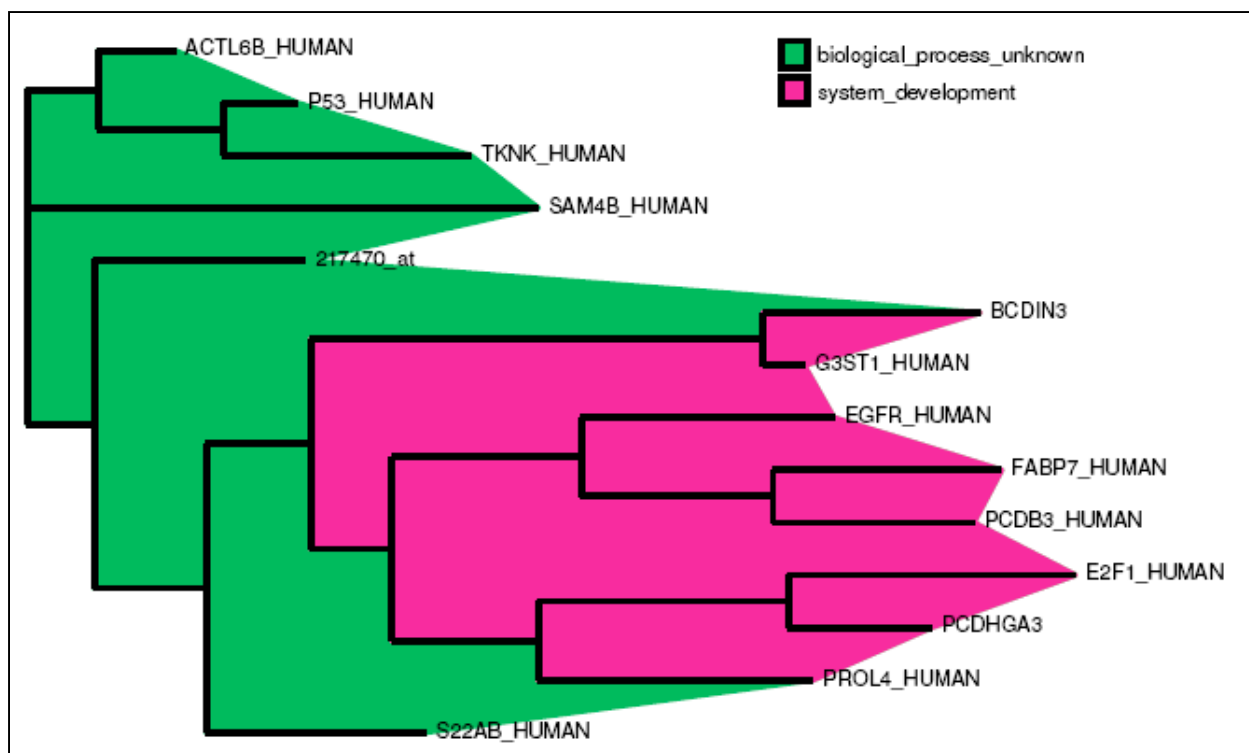
The terms cellular metabolism and protein metabolism fall into class number 1, intracellular signaling cascade, cell organization, and biogenesis fall into class number 2, and cell communication falls into class number 4. P-values are shown below in Table 4-26. The p-values are not very significant in this dataset.



**Table 4-26: p-values of Gene Ontology terms identified from known classes in Common interactions among DFCI & MSK Non-Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	6	cellular metabolism	0.3730
1	6	protein metabolism	0.3730
2	6	intracellular signaling cascade	0.3730
2	6	cell organization and biogenesis	0.3730
4	8	cell communication	0.3496

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to all the three datasets (Train data, DFCI data and the MSK data) in the Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 14 genes/proteins that were classified by the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 14 classified genes, there were 7 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 2 GO terms as shown in the Figure 4-30 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.



**Figure 4-30: Clustering from interactions Common to 3 datasets Metastasis group from PRODISTIN**

The term system development falls into class number 2. P-values are not related as shown in Table 4-27.

**Table 4-27: p-values of Gene Ontology terms identified from known classes in Common interactions of 3 datasets Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
2	8	system development	NR

The interaction network, which is a file that includes the total number of interactions and the interactions between the genes common to all the three datasets (Train data, DFCI data and the MSK data) in the Non-Metastasis group were loaded. The gene/protein connectivity was put as 1 as it is the minimal connectivity threshold. There were 18 genes/proteins that were classified by

the Prodistin Method (by computing Czekanowski-Dice distance between all possible pairs) from the uploaded network. From the 18 classified genes, there were 8 genes that were non-annotated based on the Functional class identification. The tree was drawn grouping the genes of the same functional annotation in to one class where functional annotations are derived from the GO terms. There were 4 GO terms as shown in the Figure 4-31 below. If there are multiple Gene Ontology terms for a single class, that class would be represented by a color representing one of those multiple terms. Some classes may contain other smaller classes.

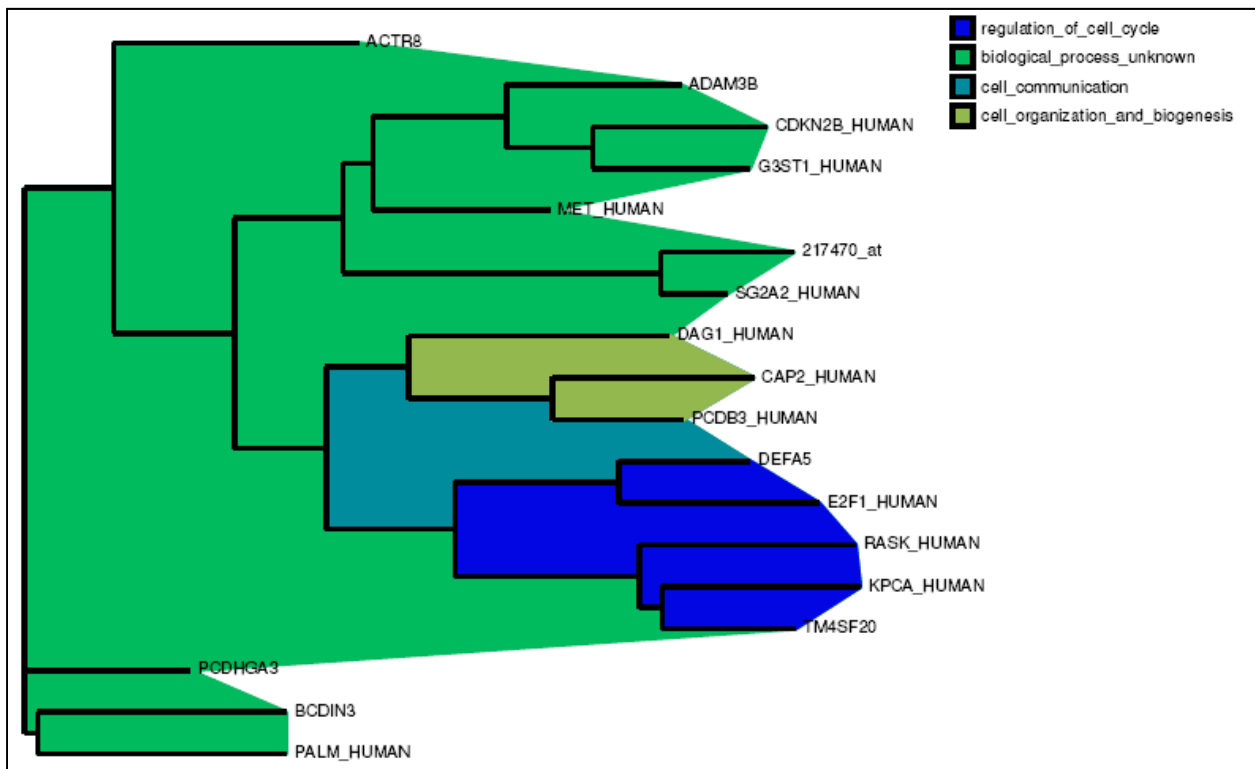


Figure 4-31: Clustering from interactions Common to 3 datasets Non-Metastasis group from PRODISTIN

The term regulation of cell cycle falls in to class number 1, cell communication falls in to class number 3, and cell organization and biogenesis falls in to class number 4. P-values are shown in

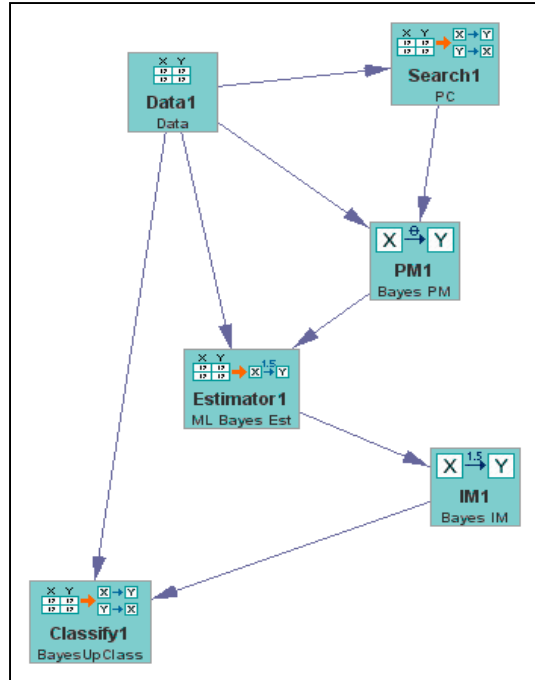
Table 4-28. The cell communication and cell organization and biogenesis classes were very significant in this dataset.

**Table 4-28: p-values of Gene Ontology terms identified from known classes in Common interactions of 3 datasets Non-Metastasis group from PRODISTIN**

Class Num	Number of genes in each class	Gene Ontology Term	P-Value
1	5	regulation of cell cycle	0.1190
3	8	cell communication	0
4	3	cell organization and biogenesis	0.0476

#### **4.5.2 TETRAD IV**

Tetrad IV is a software program used for simulating data from causal or statistical models. It is also used for estimating, testing, predicting and searching for causal or statistical models. It implements Bayes networks to generate graphical statistical/causal model for categorical data. The networks were generated using Bayesian Belief networks. The 31 genes signature were picked from all the three datasets and the data was partitioned in to 2 groups based on the survival times and status of the patients. The Metastasis group and the Non-Metastasis group were given as the data inputs to the software. The model used is shown in Figure 4-32 below.



**Figure 4-32: Model used to build Bayesian networks using Tetrad IV which uses PC search, Bayes parametric model, ML Bayes Estimator, Bayes instantiated model, and Bayes classifier**

The Metastasis groups of the datasets were given as data wrappers and PC pattern search was used on the data. DAG in pattern graph was considered as the output which was given as the input to the Parametric Model which uses Bayes parametric model. The output of the Bayes PM and the Data were given as input to the Estimator where ML Bayes Estimator is used. This output was given to the Bayes instantiated model. The output of the Bayes instantiated model along with the data is given to the Bayes updater classifier. The networks for Train Metastasis group as input is shown in Figure 4-33, DFCI Metastasis group as input is shown in Figure 4-34, and MSK Metastasis group is shown in Figure 4-35.

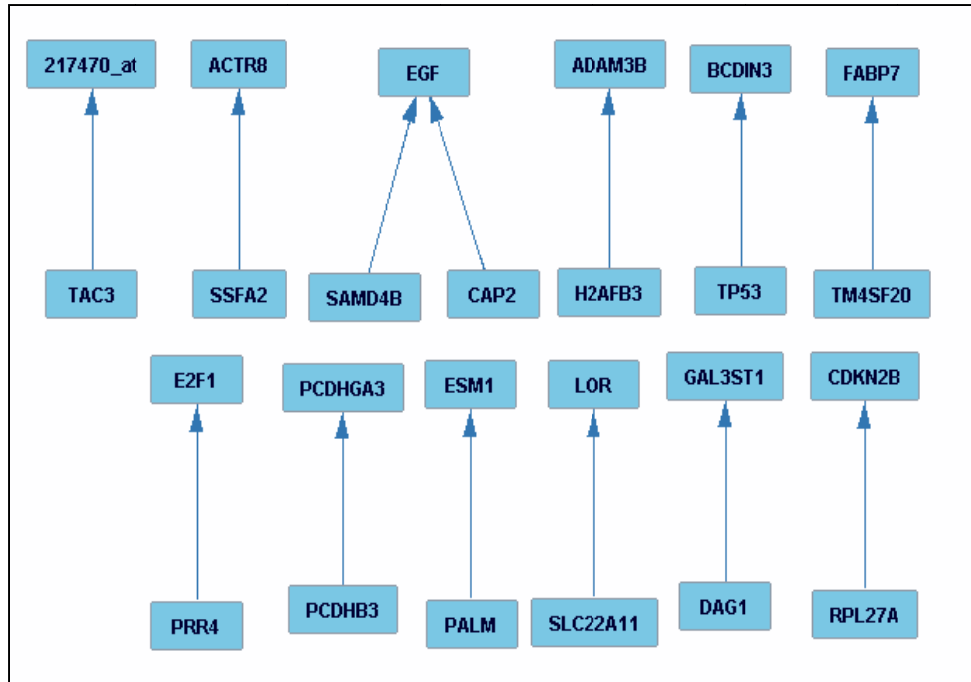


Figure 4-33: Interactions from 31 genes and the 8 hallmarks in Train Metastasis group using Tetrad IV

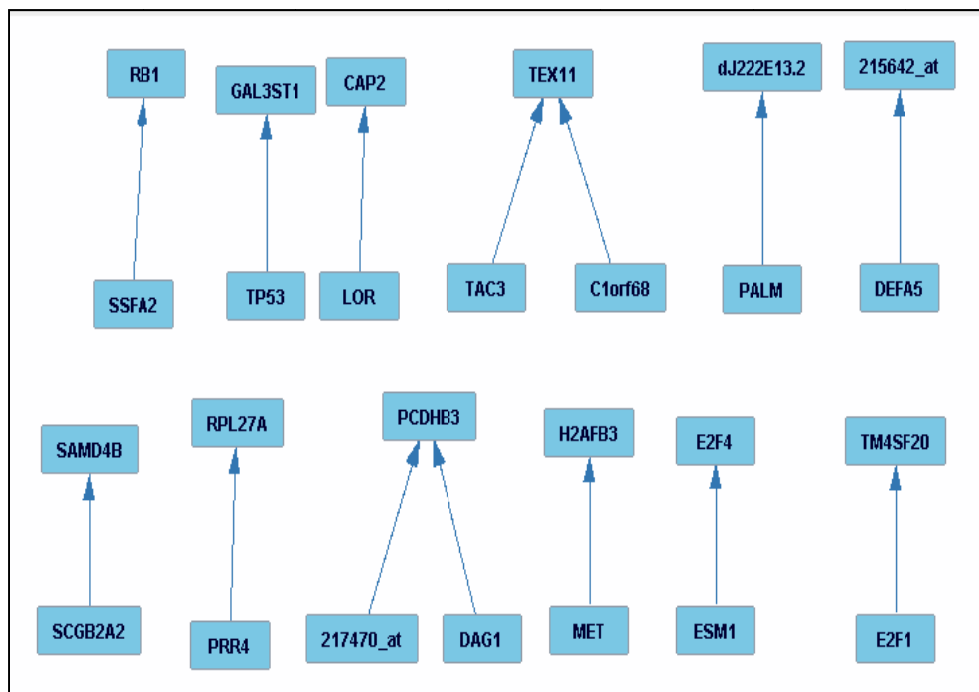


Figure 4-34: Interactions from 31 genes and the 8 hallmarks in DFCI Metastasis group using Tetrad IV

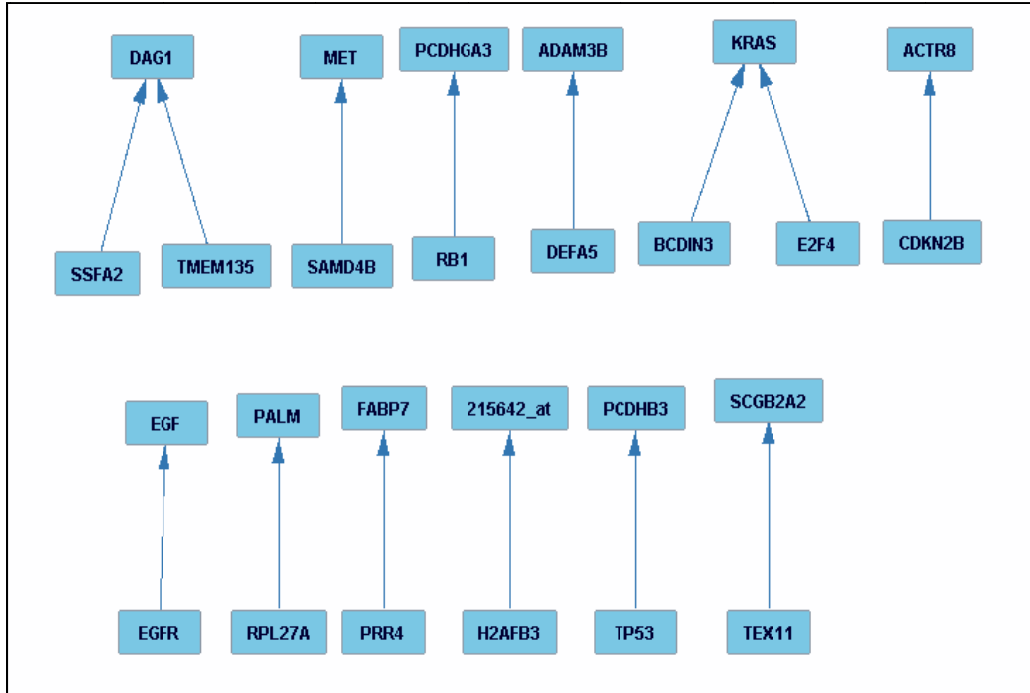


Figure 4-35: Interactions from 31 genes and the 8 hallmarks in MSK Metastasis group using Tetrad IV

The same procedure used for Metastasis group was used for Non-Metastasis group also. The only difference is that the input to Data wrappers would be Non-Metastasis group. The networks for Train Non-Metastasis group as input is shown in Figure 4-36, DFCI Non-Metastasis group as input is shown in Figure 4-37, and Non-MSK Metastasis group is shown in Figure 4-38.

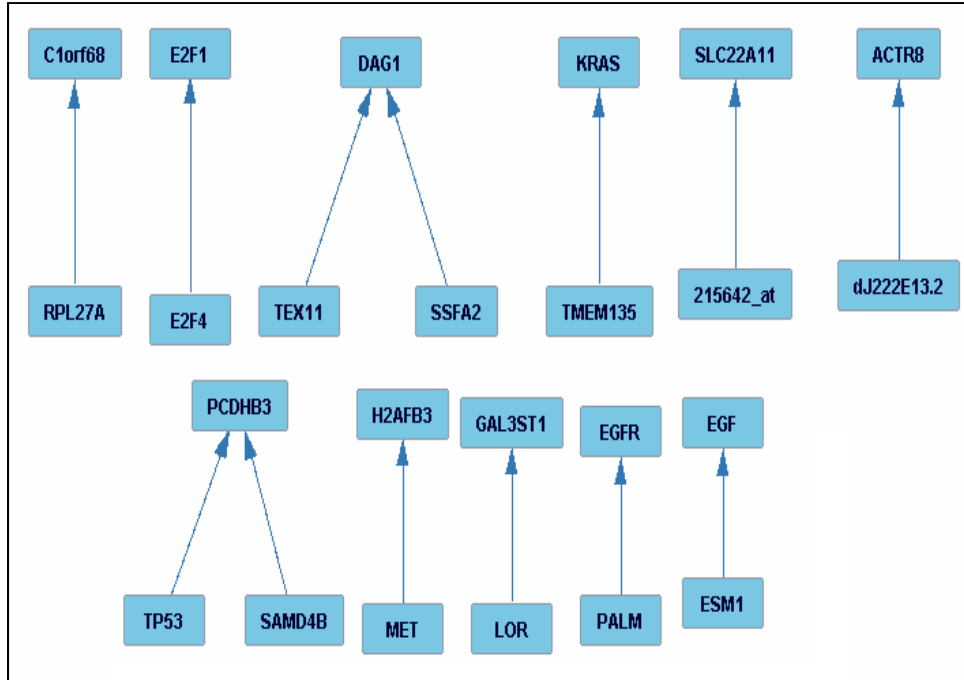


Figure 4-36: Interactions from 31 genes and the 8 hallmarks in Train Non-Metastasis group using Tetrad IV

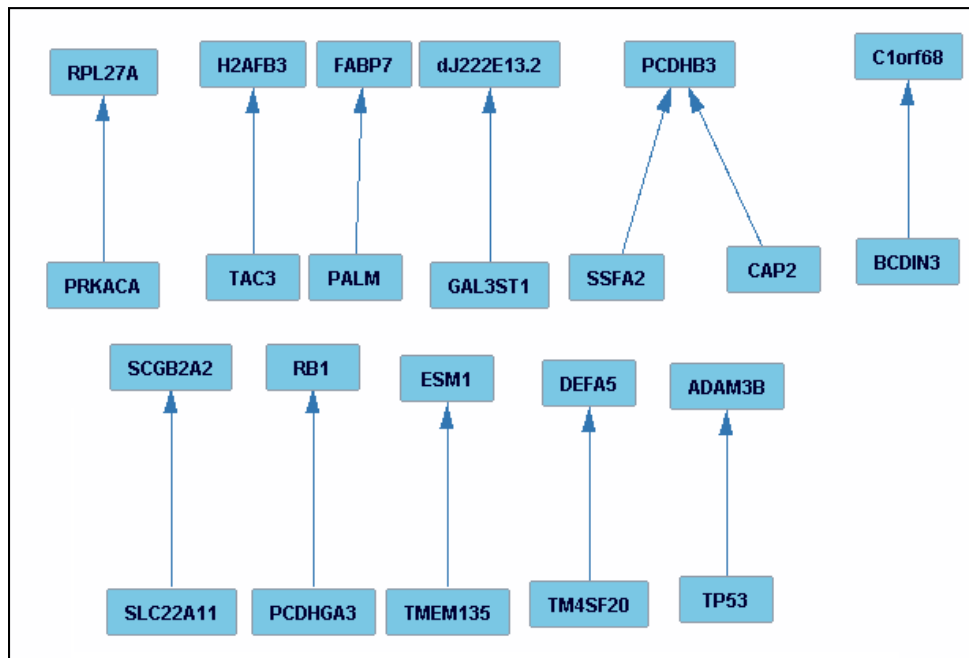
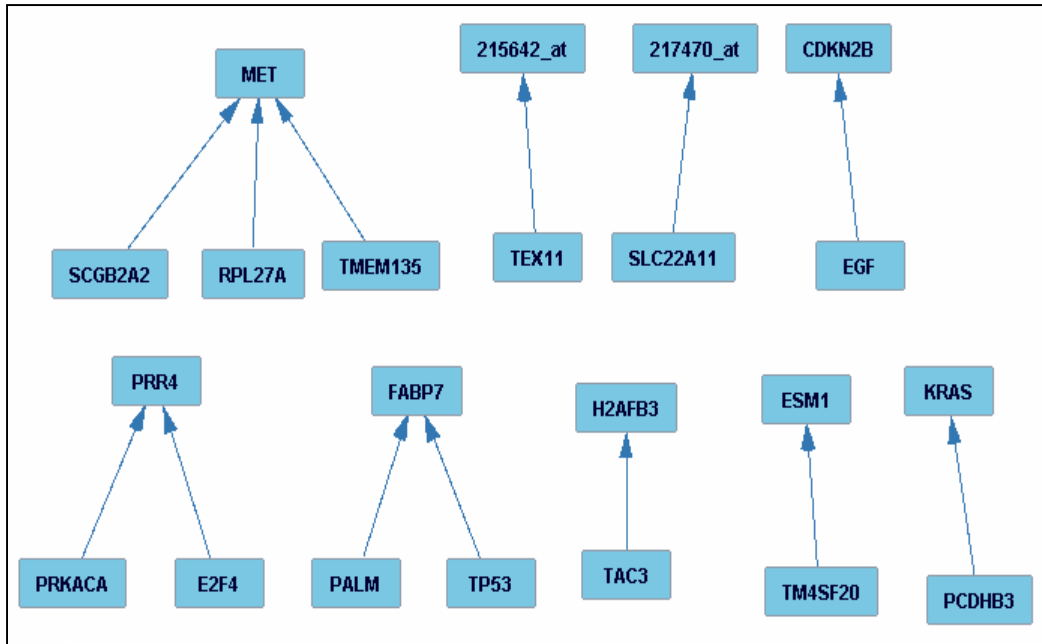


Figure 4-37: Interactions from 31 genes and the 8 hallmarks in DFCI Non-Metastasis group using Tetrad IV





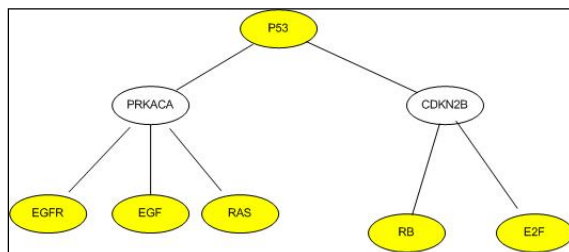
**Figure 4-38: Interactions from 31 genes and the 8 hallmarks in MSK Non-Metastasis group using Tetrad IV**

It can be seen that the 31 gene signature is connected even when using Bayesian Belief Networks. It can also be seen from the figures above that the implication networks are more connected when compared to the Bayesian Belief Networks. All the interactions from the Bayesian networks were also present in the interactions from implication networks.

### 4.5.3 KEGG

KEGG stands for Kyoto Encyclopedia for Genes and Genomes. As the name suggests, it is an encyclopedia (a large set) of genes. It is a database of 19 databases. These databases are categorized in to systems information (includes 4 databases), genomic information (includes 9 databases), and chemical information (includes 6 databases). The database that was used was the KEGG PATHWAY database which is in the systems information to find the signal pathways of the 31 genes in the signature. All the genes are searched and the genes found interacting with the remaining signature genes and hallmarks were noted. The Figure 4-39 below shows the

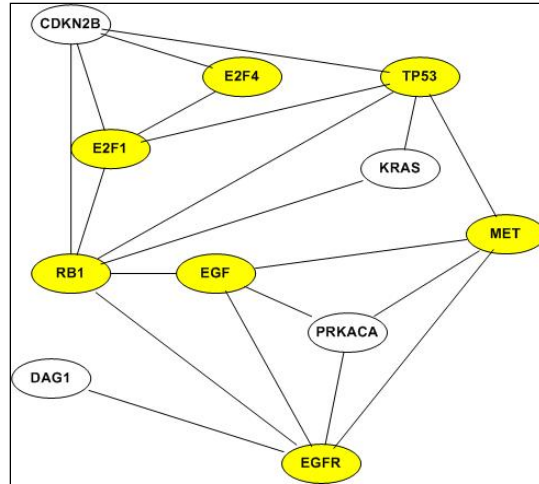
interactions that were derived from KEGG PATHWAY. The genes that are colored are the Hallmarks and the genes without colors are the signature genes. All the interactions extracted from KEGG shown below were also confirmed to be a part of the implication networks.



**Figure 4-39: Interactions among 31 genes and the 8 hallmarks extracted from KEGG PATHWAY database and all of them are confirmed with the interactions from implication networks**

#### 4.5.4 NCI Pathways

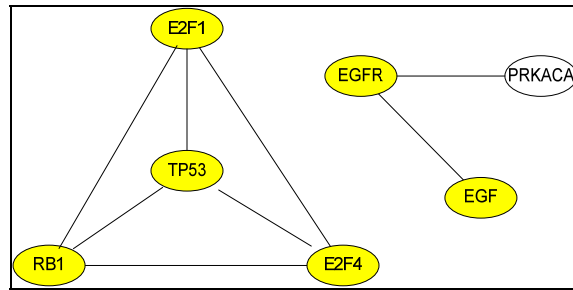
Pathway Interaction Database (PID) is a highly structured database. It is a curated collection of information about known biomolecular interactions and key cellular processes assembled in to authoritative human signaling pathways. It includes pathways from various reliable sources such as NCI-Nature curated data, BioCarta data, and Reactome data. All the signature genes and the hallmarks were searched in the pathways and those genes found to be interacting with one another were noted down. The Figure 4-40 below shows the interactions that were derived from PID. The genes that are colored are the Hallmarks and the genes without colors are the signature genes. All the interactions extracted from NCI shown below were also confirmed to be a part of the implication networks.



**Figure 4-40: Interactions among 31 genes and the 8 hallmarks extracted from NCI pathways and all of them are confirmed with the interactions from implication networks**

#### 4.5.5 PubMed interactions

PubMed was developed at the National Library of Medicine (NLM) which was located at the US National Institutes of Health (NIH). It was developed by National Center for Biotechnology Information (NCBI). It is a search engine which includes accesses to many databases in the field of medicine and related disciplines. It also holds the links to an enormous number of citations, abstracts, journals and full text articles. The signature genes and the hallmarks were searched in PubMed and their interactors were noted if they were present among the signature genes or the hallmarks. The Figure 4-41 below shows the interactions between the signature genes and the Hallmarks. The genes that are colored are the Hallmarks and the genes without colors are the signature genes. All the interactions extracted from PubMed shown below were also confirmed to be a part of the implication networks.



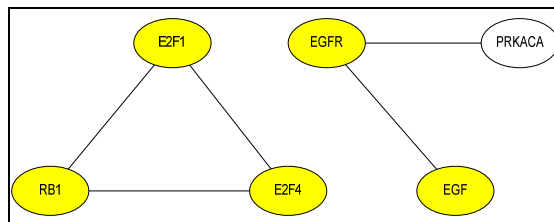
**Figure 4-41: Interactions among 31 genes and the 8 hallmarks extracted from PubMed and all of them are confirmed with the interactions from implication networks**

#### 4.5.6 Matisse

Matisse stands for Modular Analysis via Topology of Interactions and Similarity Sets. It is a software program for detecting the functional modules present in a set of data. It uses an interaction network which has already been generated from trustworthy sources. It acts as a tool to identify sets of genes that are highly correlated and also connected sub graphs in networks.

The species was selected as Homo sapiens. An interaction network for human genomes (pre generated) was loaded. The gene expression file of the Training dataset for the 31 gene signature along with the hallmarks was loaded and the program was ran which detects the nodes and edges (6214 and 25086 respectively) of the interaction network and expression patterns and conditions (39 and 256) of the dataset loaded. New Modules were found using different algorithms like Matisse and Expression k-means. The minimum seed and module sizes were varied between 1 and 5. Correlation coefficients were found using one of the various methods such as Dot product (Pearson), Euclidean distance, Spearman correlation, and Partial correlation. The modules contained many genes which were not present in the 31 gene signature and so they were ignored. The gene interactions which included the genes from the signature and Hallmarks are shown in

the Figure 4-42 below. All the interactions extracted from Matisse shown below were also confirmed to be a part of the implication networks.

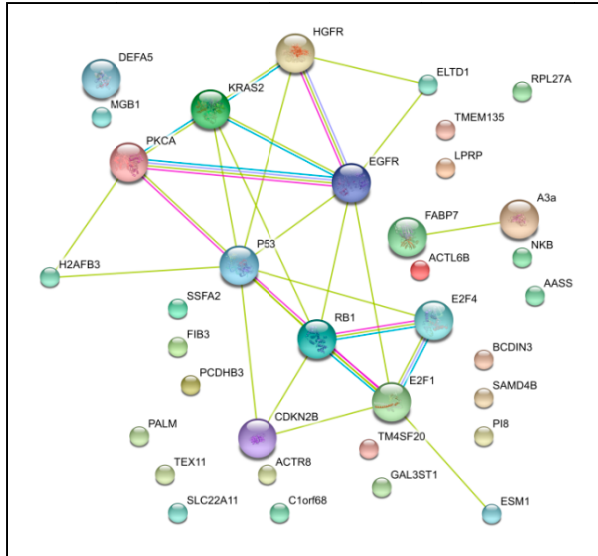


**Figure 4-42: Interactions among 31 genes and the 8 hallmarks extracted from MATISSE and all of them are confirmed with the interactions from implication networks**

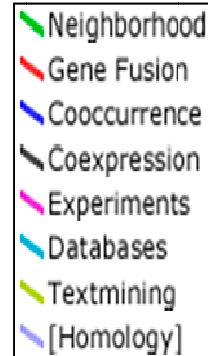
#### 4.5.7 STRING 8

STRING stands for Search Tool for the Retrieval of Interacting Genes/Proteins. It is web based tool used to extract the protein-protein interactions between the set of genes that were input to it. It also includes interactions from various other sources such as MINT, HPRD, BIND, DIP, etc. other than the interactions that were extracted from its algorithm.

The 31 genes along with the 8 hallmarks were input to STRING. It identified all the genes from its database in various species and generates a list where the most probable species was highlighted at the top of the list. Once the species was selected, it gave a list of aliases for each of the genes with the most important one highlighted. Some genes might not be found in its database. After the required genes were selected, it generated a figure of the network that was generated using medium confidence of 0.4 as a default value. The evidence view of the network generated is shown in the Figure 4-43 below. Each color in the interactions corresponds to a different source as shown in Figure 4-44. All the interactions extracted from STRING shown below were confirmed to be present in the implication networks.



**Figure 4-43: Evidence view of the interactions between the 31 genes and the 8 hallmarks and all of them were confirmed with the interactions from implication networks**



**Figure 4-44: Various sources of identification of gene interactions in STRING 8**

String also gives the list of genes that have been identified along with their sources at the bottom of the network as shown in Figure 4-45 below. The color of the bullet beside the gene name gives its source.

Your Input:	
● ACTL6B	Actin-like protein 6B (53 kDa BRG1-associated factor B) (Actin-related protein Baf53b) (ArpNalpha) (426 aa)
● LPRP	Proline-rich protein 4 precursor (Lacrimal proline-rich protein) (Nasopharyngeal carcinoma-associated proline-rich protein 4) (134 aa)
● PCDHB3	Protocadherin beta 3 precursor (PCDH-beta3) (796 aa)
● FIB3	Protocadherin gamma A12 precursor (PCDH-gamma-A12) (Cadherin-21) (Fibroblast cadherin 3) (944 aa)
● KRAS2	GTPase KRas precursor (K-Ras 2) (Ki-Ras) (c-K-ras) (c-Ki-ras) (189 aa)
● SSFA2	Sperm-specific antigen 2 (Cleavage signal-1 protein) (CS-1) (Ki-ras- induced actin-interacting protein) (1259 aa)
● RB1	Retinoblastoma-associated protein (P110) (P105-RB) (RB) (928 aa)
● P53	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13) (393 aa)
● EGFR	Epidermal growth factor receptor precursor (EC 2.7.10.1) (Receptor tyrosine-protein kinase ErbB-1) (1210 aa)
● CDKN2B	Cyclin-dependent kinase 4 inhibitor B (p14-INK4b) (p15-INK4b) (p15INK4B) (Multiple tumor suppressor 2) (MTS2) (138 aa)
● PKCA	Protein kinase C alpha type (EC 2.7.11.13) (PKC-alpha) (PKC-A) (672 aa)
● TM4SF20	Transmembrane 4 L6 family member 20 (229 aa)
● TMEM135	Transmembrane protein 135 (458 aa)
● BCDIN3	bin3, bicoid-interacting 3 (689 aa)
● A3a	Dystroglycan precursor (Dystroglycan-associated glycoprotein 1) [Contains- Alpha-dystroglycan (Alpha-DG); Beta-dystroglycan (Beta- DG)] (895 aa)
● SAMD4B	Sterile alpha motif domain-containing protein 4B (694 aa)
● HGFR	Hepatocyte growth factor receptor precursor (EC 2.7.10.1) (HGF receptor) (Scatter factor receptor) (SF receptor) (HGF/SF receptor) (Met proto-oncogene tyrosine kinase) (c-Met) (1408 aa)
● PI8	Serpin B8 (Cytoplasmic antiproteinase 2) (CAP-2) (CAP2) (Protease inhibitor 8) (374 aa)
● ACTR8	Actin-related protein 8 (624 aa)
● TEX11	testis expressed sequence 11 isoform 1 (940 aa)
● PALM	Paralemmin (387 aa)
● GAL3ST1	Galactosylceramide sulfotransferase (EC 2.8.2.11) (GalCer sulfotransferase) (Cerebroside sulfotransferase) (3'- phosphoadenylylsulfate-galactosylceramide 3'-sulfotransferase) (3'- phosphoadenosine-5'-phosphosulfate-GalCer sulfotransferase) (423 aa)
● E2F1	Transcription factor E2F1 (E2F-1) (Retinoblastoma-binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (FBR3) (Retinoblastoma-associated protein 1) (RBAP-1) (437 aa)
● RPL27A	60S ribosomal protein L27a (148 aa)
● FABP7	Fatty acid-binding protein, brain (B-FABP) (Brain lipid-binding protein) (BLBP) (Mammary-derived growth inhibitor related) (166 aa)
● NKB	Neurokinin-B precursor (NKB) (Neuromedin-K) (Tachykinin-3) (ZNEUROK1) (135 aa)
● AASS	Alpha-aminoacidic semialdehyde synthase, mitochondrial precursor (LKR/SDH) [Includes- Lysine ketoglutarate reductase (EC 1.5.1.8) (LOR) (LKR); Saccharopine dehydrogenase (EC 1.5.1.9) (SDH)] (926 aa)
● C1orf68	Uncharacterized protein C1orf68 (Late envelope protein 7) (Skin- specific protein xp32) (250 aa)
● H2AFB3	Histone H2A-Bbd (H2A Barr body-deficient) (H2A.Bbd) (115 aa)
● ELTD1	EGF, latrophilin and seven transmembrane domain-containing protein 1 precursor (EGF-TM7-latrophilin-related protein) (ETL protein) (690 aa)
● SLC22A11	Solute carrier family 22 member 11 (Organic anion transporter 4) (550 aa)
● MGB1	Mammaglobin-A precursor (Mammaglobin-1) (Secretoglobin family 2A member 2) (120 aa)
● E2F4	Transcription factor E2F4 (E2F-4) (417 aa)
● ESM1	Endothelial cell-specific molecule 1 precursor (ESM-1 secretory protein) (ESM-1) (184 aa)
● DEFA5	Defensin 5 precursor (Defensin, alpha 5) (94 aa)

**Figure 4-45: List of the input genes (among 31 gene signature) that were identified by STRING 8 and were displayed at the output**

#### 4.5.8 Ingenuity Pathway Analysis

When the 31 genes along with the 8 hallmarks were input in to IPA, it generated five networks that were significant. The networks contained not only the genes from the signature but also those that that played an important role in the network. The five networks are shown below in the Figures 4-46 to 4-50.

The first network shown in Figure 4-46 had 33 molecules and is associated with Network Functions such as Cancer, Cellular Growth and Proliferation, and Hematological Disease. The second network shown in Figure 4-47 is associated with Network Functions such as Cancer, Cell Cycle, and Cellular Development. The third network shown in Figure 4-48 is associated with Network Functions such as Cell Morphology, Cellular Assembly and Organization, and

Connective Tissue Development and Function. The fourth network shown in Figure 4-49 is associated with Network Functions such as Infection Mechanism, Cancer, and Hepatic System Disease. The fifth network shown in Figure 4-50 is associated with Network Functions such as Infection Mechanism, Gene Expression, and Cancer.

There are a lot of Bio Functions under the Diseases and Disorder, Molecular and Cellular Functions, Physiological System Development and Function that were found from the signature which had very significant p-values. There were also Canonical Pathways, Tox lists, and Tox Functions which were significant.

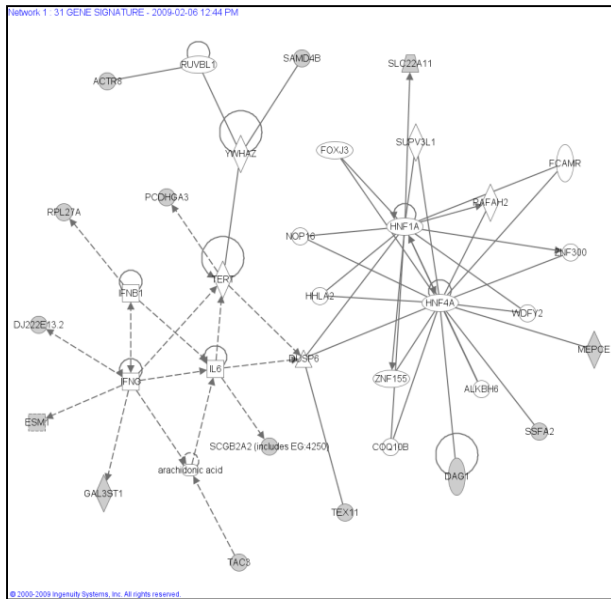


Figure 4-46: Network 1 generated from IPA

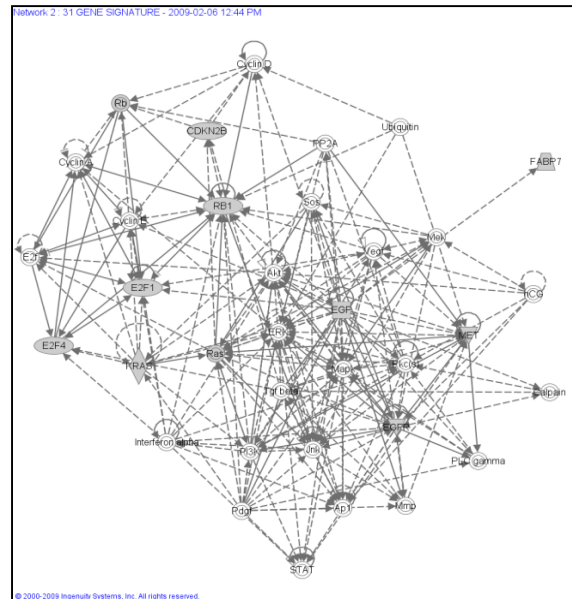
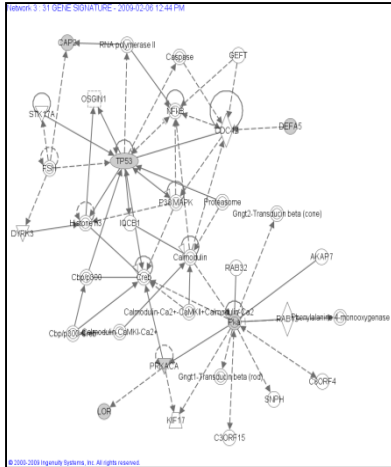
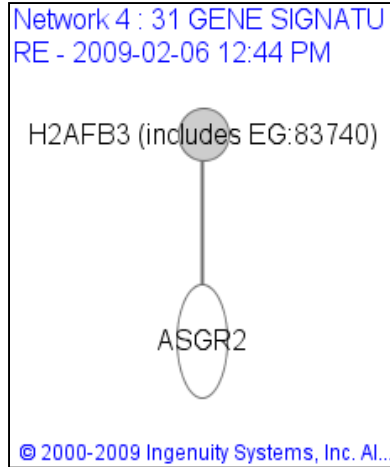


Figure 4-47: Network 2 generated from IPA

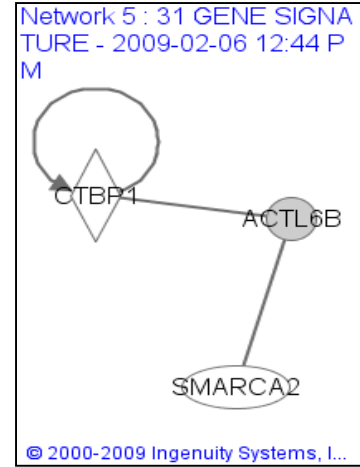




**Figure 4-48: Network 3 generated from IPA**

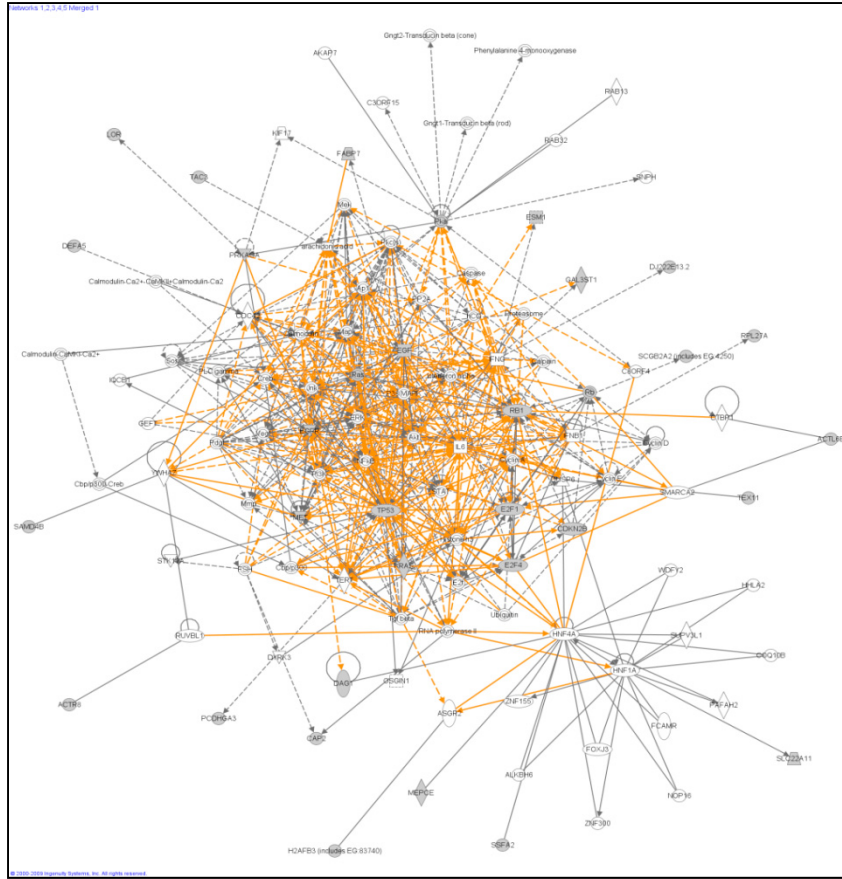


**Figure 4-49: Network 4 generated from IPA**



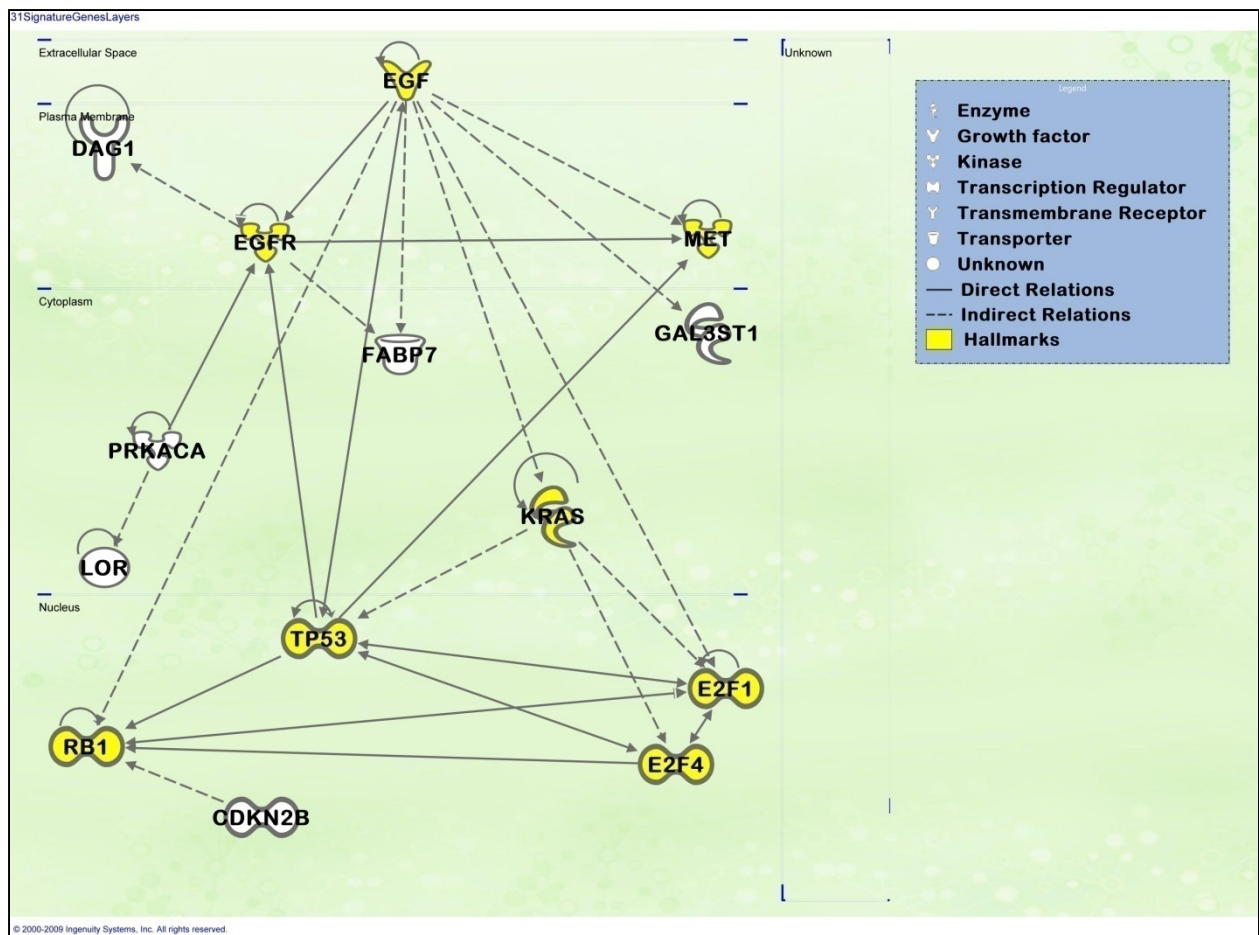
**Figure 4-50: Network 5 generated from IPA**

When all the five networks were merged to form a big network, it had two types of connections. The highlighted interactions between the genes were the inter network connections that did not exist in the five networks shown above. They were emerged just because of the merging. This merged network is shown below in Figure 4-51.



**Figure 4-51: Merged network from all the 5 networks shown above where grey connections are the intra-network connections and orange connections are inter-network connections in IPA**

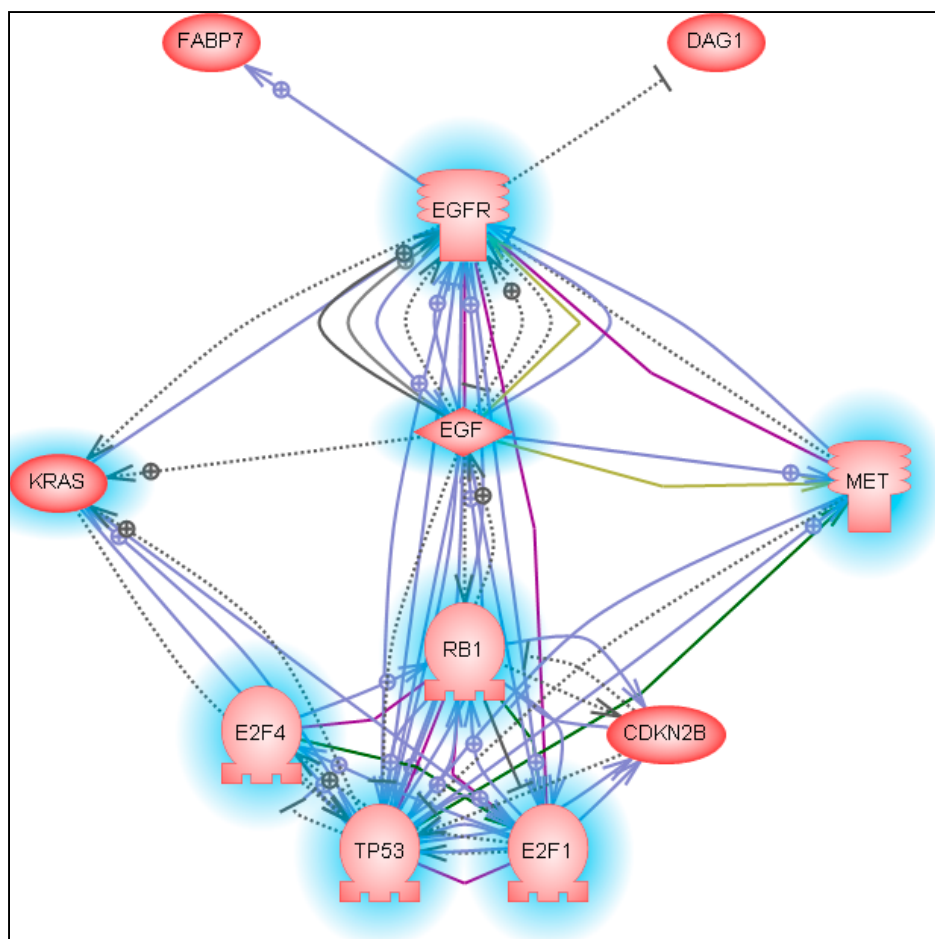
All the interactions involving the 31 genes signature and the 8 hallmarks only were separated. These interactions were confirmed to be present in the implication network. These interactions are shown in the Figure 4-52 below. Each node is of different shape and the legend explains the meaning of each of the shapes. The solid lines represent the direct interactions and the dotted lines represent the indirect interactions.



**Figure 4-52: Interactions between the 31 genes and the 8 cancer hallmarks extracted from the merged network of all the 5 networks in IPA shown above where the yellow genes represent the major cancer hallmarks**

### 4.5.9 Pathway Studio

The 31 genes signature and the 8 hallmarks were input to Pathway Studio. All the genes except 215642\_at and 217470\_at were found. Hence the signature contained only 29 genes which were fed into Pathway Studio. It generated a network as shown below in Figure 4-61. There were numerous interactions in between a pair of genes which indicate that different kinds of relationships were found between those genes from different sources. All the interactions shown in the Figure below were confirmed to be present in the implication network.



**Figure 4-53: Interactions between the 31 genes and the 8 hallmarks that were extracted from Pathway studio where each kind of line represents different kinds of relationships between the genes**

## 4.6 Summary

This chapter provided the results that were obtained from the performed analyses. Thus the number of interactions of implication networks and Bayesian networks, in different datasets for the 31 gene signature in each of the groups are concluded below in Table 4-29. It can be seen that the implication networks were able to detect many more gene/protein interactions when compared to the Bayesian networks.

**Table 4-29: Comparison of number of interactions from Poor and Good Prognosis of each dataset generated in Implication Networks and Bayesian Networks (using Tetrad IV)**

	IMPLICATION NETWORKS	TETRAD IV (BAYESIAN)
TRAINING GOOD PROGNOSIS	897	13
TRAINING POOR PROGNOSIS	1021	13
DFCI GOOD PROGNOSIS	938	12
DFCI POOR PROGNOSIS	787	14
MSK GOOD PROGNOSIS	996	13
MSK POOR PROGNOSIS	992	14

The interactions among the 31 genes extracted from different tools were compared with the interactions obtained from implication networks and Bayesian networks as concluded in Table 4-30. It can be seen that the interactions from all the tools were present in implication networks but most of them did not show up in Bayesian networks.

**Table 4-30: Comparison of number of interactions among the 31 genes and the 8 hallmarks identified from different biomedical tools found in implication networks and Bayesian networks**

	IMPLICATION NETWORKS	TETRAD IV (BAYESIAN)
MATISSE (8)	100% (8/8)	12.5% (1/8)
PUBMED (5)	100% (5/5)	20% (1/5)
KEGG (7)	100% (7/7)	0% (0/7)
NCI (20)	100% (20/20)	0% (0/20)
STRING(27)	100% (27/27)	3.7% (1/27)
PATHWAY STUDIO (26)	100% (26/26)	3.84% (1/26)
INGENUITY PATHWAY (24)	100% (24/24)	4.16% (1/24)

The interactions from implication networks and Bayesian networks were input in to Prodistin and the biological processes they are involved are noted down along with the number of significant processes. It can be seen from Table 4-31 that interactions from Bayesian networks did not show any biological processes. On the other hand, interaction from implication networks consisted of many biological processes.

**Table 4-31: Number of Biological Processes identified using Prodistin when interactions from implication networks and Bayesian networks are given as input.**

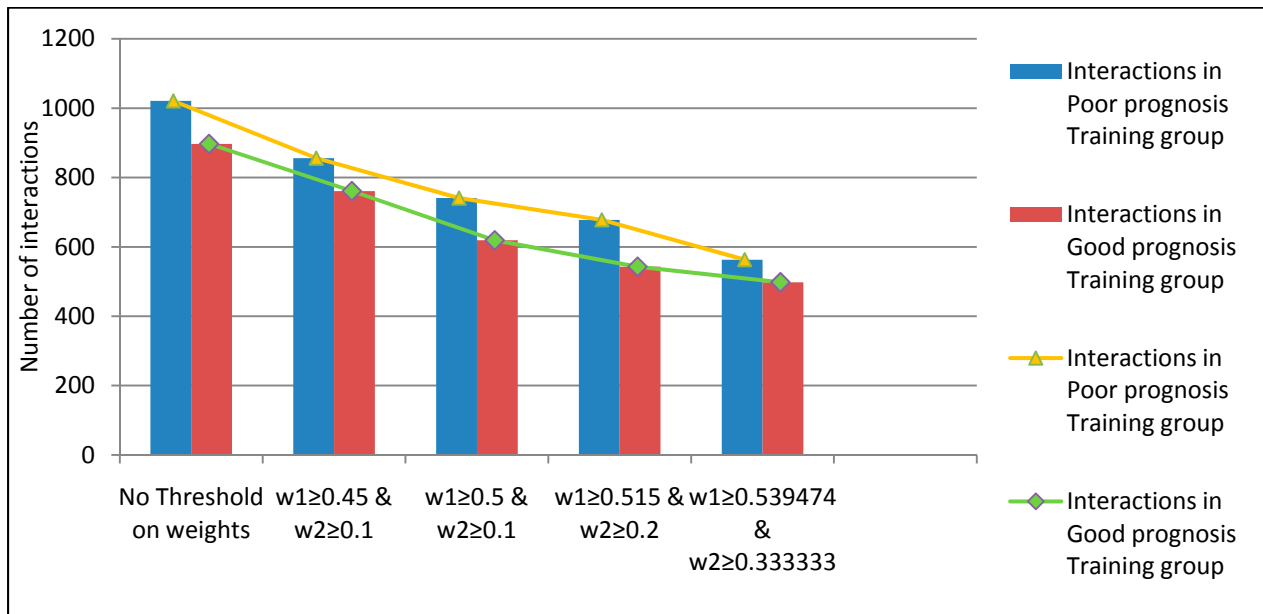
	IMPLICATION NETWORKS	TETRAD IV (BAYESIAN)
TRAINING GOOD PROGNOSIS	14(3 SIGNIFICANT)	0
TRAINING POOR PROGNOSIS	7(2 SIGNIFICANT)	0
DFCI GOOD PROGNOSIS	4	0
DFCI POOR PROGNOSIS	2	0
MSK GOOD PROGNOSIS	11 (3 SIGNIFICANT)	0
MSK POOR PROGNOSIS	11	0

To increase the possibility of biological relevance and to reduce the false discovery rate, thresholds can be applied on the weight functions of the implication rules. For the results in Table 4-29, no thresholds were applied and hence there were a large number of interactions. We applied a threshold on the weight functions (equations 2.17 and 2.18) to reduce the number of interactions that would remain along with all the interactions from curated databases. The weights are variables between [0, 1]. The thresholds  $w_1 \geq 0.539474$  and  $w_2 \geq 0.333333$  made all the interactions from curated databases remain in the implication networks. For these applied thresholds, the number of interactions that remain in each of the datasets in Poor and Good prognosis are summarized in the Table 4-32 below. A few intermediate calculations of threshold on weights are also shown in the table below.

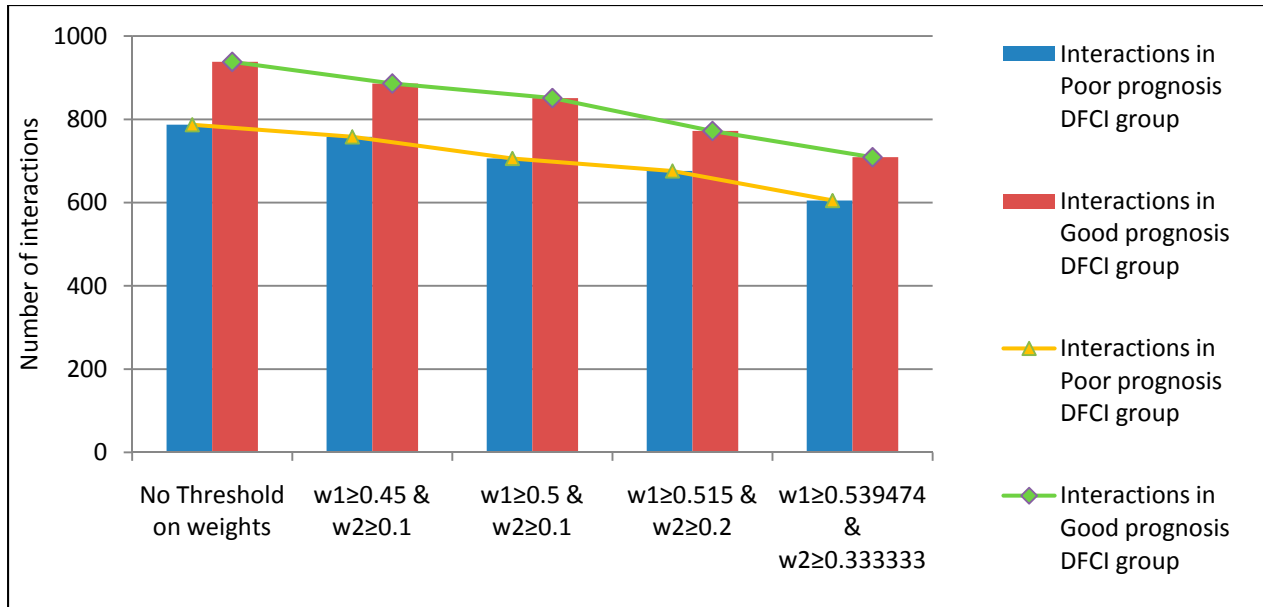
**Table 4-32: Comparison of number of interactions from Poor and Good Prognosis of each dataset in Implication Networks and Bayesian Networks (using Tetrad IV) with application of thresholds on weights.**

	No Threshold on weights	$w_1 \geq 0.45$ & $w_2 \geq 0.1$	$w_1 \geq 0.5$ & $w_2 \geq 0.1$	$w_1 \geq 0.515$ & $w_2 \geq 0.2$	$w_1 \geq 0.539$ & $w_2 \geq 0.333$
TRAINING GOOD PROGNOSIS	897	761	619	543	498
TRAINING POOR PROGNOSIS	1021	856	741	678	563
DFCI GOOD PROGNOSIS	938	886	851	772	709
DFCI POOR PROGNOSIS	787	758	706	676	605
MSK GOOD PROGNOSIS	996	949	911	849	752
MSK POOR PROGNOSIS	992	935	877	817	742

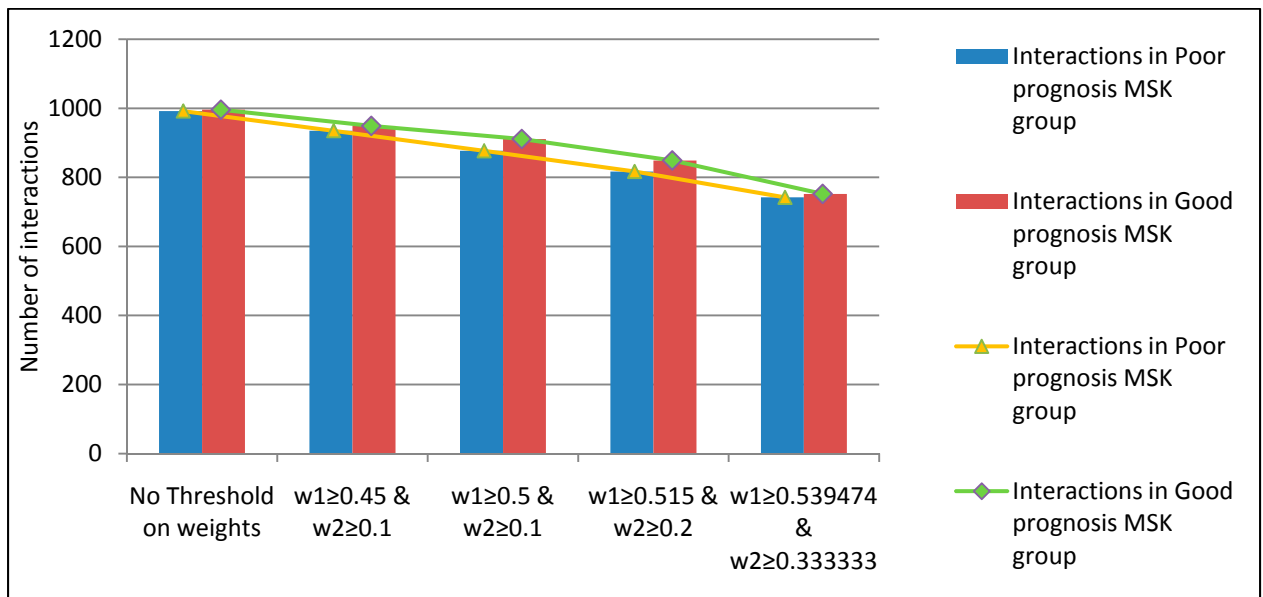
The Figures below show the representation of the values in the Table above for the three datasets separately. Figure 4-54 shows the variation of the number of gene interactions with threshold on weights in the Training group. Figure 4-55 shows the variation of the number of gene interactions with threshold on weights in the DFCI test group. Figure 4-56 shows the variation of the number of gene interactions with threshold on weights in the MSK test group.



**Figure 4-54: Variation of number of gene interactions with threshold on weights in Training Group.** The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds.



**Figure 4-55: Variation of number of gene interactions with threshold on weights in DFCI test Group.** The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds.



**Figure 4-56: Variation of number of gene interactions with threshold on weights in MSK test Group.** The first set of data is the number of interactions without any thresholds and the fifth set of data is the number of gene interactions with the given thresholds which include all the curated interactions. The second, third and fourth set of data are intermediate set of results to show how the number of gene interactions decrease with an increase in thresholds.



This chapter thus summarizes that the 31 gene signature that has been identified using implication networks and interactions with hallmarks is a good predictor. It also summarizes that the model used for generating the 31 gene signature is also very good in detecting more gene/protein signatures when compared with Bayesian networks.

## **5 Software Implementation**

### **5.1. Introduction**

This chapter describes the implementation of the package which was used to perform the analyses. The package is a combination of C and R where C-code was made to run through the R-interface. This chapter also describes the different versions of the code available and the changes between them. It also gives a few screen shots of the implementation of the package and the configuration of the computer used to run the code.

### **5.2. Description**

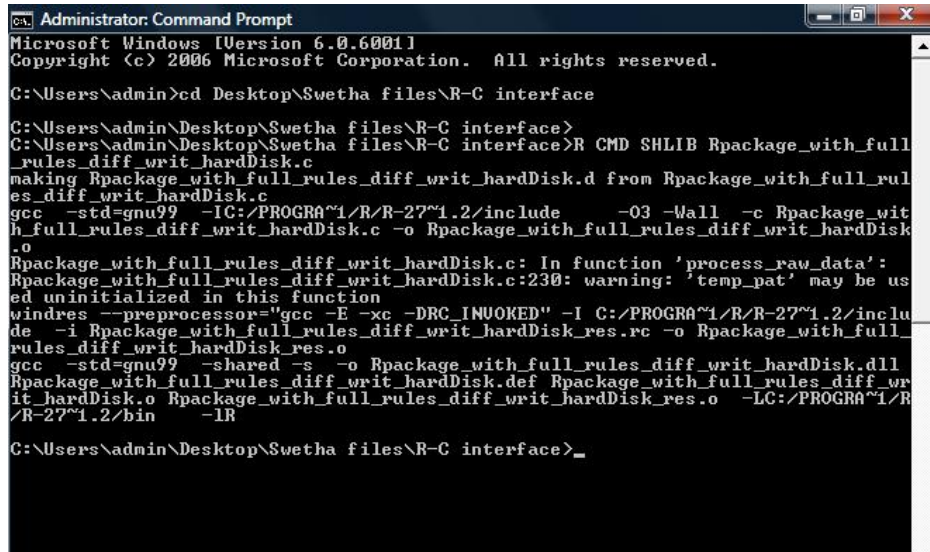
All the code was implemented in C. It was compiled at the command prompt to create the .dll (dynamic linked library) file. After the compilation, the .dll file was loaded in to R. In R, the code needed 4 filenames as input. Two of these filenames correspond to the files that were given as input while the other two file names were the names of the output files that were created while the code executed and the final output was stored. After the execution was over the .dll file that was loaded initially had to be unloaded to avoid errors in the later executions. The first of the input files contains the micro array data of patients profile gene expression values of all the genes along with their gene symbols and the survival time and status at the end of the file. The second file is the list of the hallmarks that were used. The two output files contain the genes that interact with all the hallmarks, one file for each of the groups.

There are two versions of the C-codes. Both of them work in the above mentioned manner. But the main difference between them is the speed of execution and the amount of memory utilized.

First version of the code generates the entire genome wide interactions and keeps storing them in the hard disk of the computer while generating. It takes time as there are a lot of memory read and write operations. The second version of the code generates only the interactions between the genes and the hallmarks ignoring all the interactions that do not contain at least one of the hallmarks. This code does not require more memory as it uses linked lists and stores the interactions in the cache. Thus it is much faster than the first version.

It required around 40 minutes for executing the first version of the package through R. The second version required around 25 minutes. The codes were executed on a system with the following configuration: The processor was an Intel® Core™2 Duo CPU E8300 @ 2.83GHz. There was 4.00GB of Memory (RAM) in the system. The C drive was allocated with 455GB of hard disk space. The version of the R editor used was R-2.7.2 and the C editor used was Dev-C++ 4.9.9.2.

### 5.3. Results & Screenshots

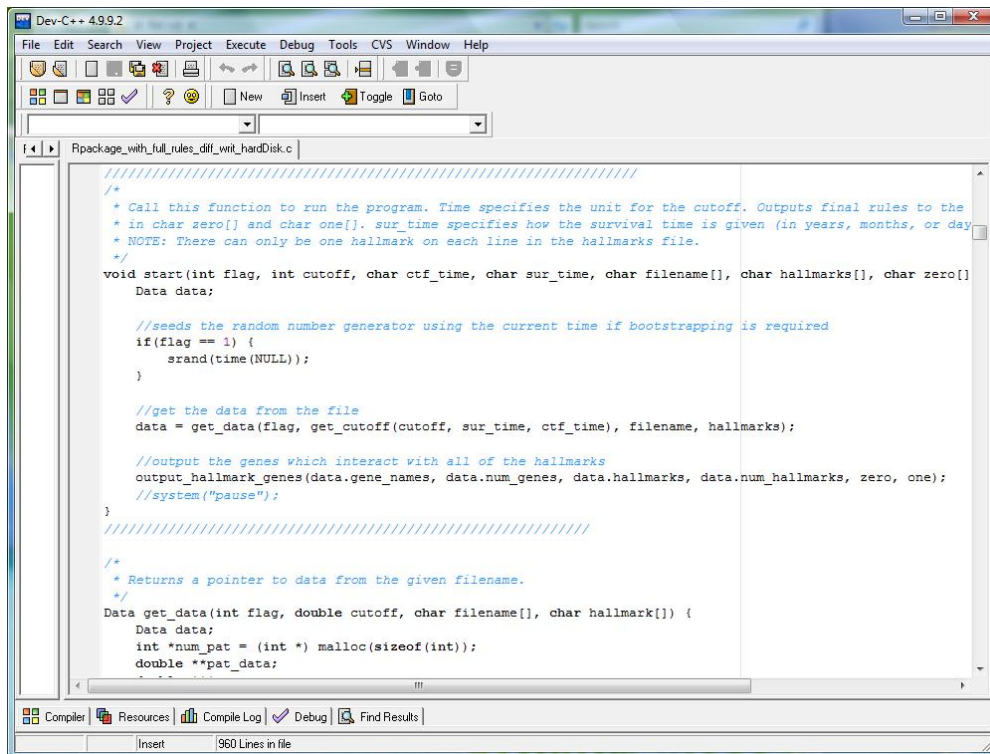


```
Administrator: Command Prompt
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation. All rights reserved.

C:\Users\admin>cd Desktop\Swetha files\R-C interface
C:\Users\admin\Desktop\Swetha files\R-C interface>
C:\Users\admin\Desktop\Swetha files\R-C interface>R CMD SHLIB Rpackage_with_full_rules_diff_writ_hardDisk.c
making Rpackage_with_full_rules_diff_writ_hardDisk.d from Rpackage_with_full_rules_diff_writ_hardDisk.c
gcc -std=gnu99 -IC:/PROGRAM~1/R/R-27~1.2/include -O3 -Wall -c Rpackage_with_full_rules_diff_writ_hardDisk.c -o Rpackage_with_full_rules_diff_writ_hardDisk.o
Rpackage_with_full_rules_diff_writ_hardDisk.c: In function 'process_raw_data':
Rpackage_with_full_rules_diff_writ_hardDisk.c:230: warning: 'temp_pat' may be used uninitialized in this function
windres --preprocessor="gcc -E -xc -DRC_INUOKED" -I C:/PROGRAM~1/R/R-27~1.2/include -i Rpackage_with_full_rules_diff_writ_hardDisk_res.rc -o Rpackage_with_full_rules_diff_writ_hardDisk_res.o
gcc -std=gnu99 -shared -s -o Rpackage_with_full_rules_diff_writ_hardDisk.dll Rpackage_with_full_rules_diff_writ_hardDisk.def Rpackage_with_full_rules_diff_writ_hardDisk.o Rpackage_with_full_rules_diff_writ_hardDisk_res.o -LC:/PROGRAM~1/R/R-27~1.2/bin -lR

C:\Users\admin\Desktop\Swetha files\R-C interface>_
```

Figure 5-1: Changing the directory to the current directory and compiling the C-code to generate the required dynamic linked library files to be used for executing code in R



```
Dev-C++ 4.9.9.2
File Edit Search View Project Execute Debug Tools CVS Window Help
Rpackage_with_full_rules_diff_writ_hardDisk.c
/*
 * Call this function to run the program. Time specifies the unit for the cutoff. Outputs final rules to the
 * in char zero[] and char one[]. sur_time specifies how the survival time is given (in years, months, or day
 * NOTE: There can only be one hallmark on each line in the hallmarks file.
 */
void start(int flag, int cutoff, char ctf_time, char sur_time, char filename[], char hallmarks[], char zero[]
Data data;

//seeds the random number generator using the current time if bootstrapping is required
if(flag == 1) {
    srand(time(NULL));
}

//get the data from the file
data = get_data(flag, get_cutoff(cutoff, sur_time, ctf_time), filename, hallmarks);

//output the genes which interact with all of the hallmarks
output_hallmark_genes(data.gene_names, data.num_genes, data.hallmarks, data.num_hallmarks, zero, one);
//system("pause");
}

/*
 * Returns a pointer to data from the given filename.
 */
Data get_data(int flag, double cutoff, char filename[], char hallmark[]) {
Data data;
int *num_pat = (int *) malloc(sizeof(int));
double **pat_data;
```

Figure 5-2: C-code for the first version of the package

```

/*
 * Generate the rules from the given data.
 */
Rule * generate_rules(double min_x, int **data, int cols, int rows, int *hallmarks, int num_hallmarks) {
    int i, j, k;
    double max_u, del1, del2, del3, del4, del5, del6;
    double u_1, u_2, u_3, u_4, u_5, u_6;
    int op, count_ab, count_anb, count_nab, count_nanb, once;
    Rule *temp = NULL, *rules;

    for(i = 0; i < rows - 1; ++i) {
        for(j = i + 1; j < rows; ++j) {
            //skip this gene neither i nor j are hallmarks or both i and j are hallmarks
            if(contains(hallmarks, num_hallmarks, i) == contains(hallmarks, num_hallmarks, j)) {
                continue;
            }
            max_u = min_x;
            del1 = 0.0; del2 = 0.0; del3 = 0.0; del4 = 0.0; del5 = 0.0; del6 = 0.0;
            u_1 = 0.0; u_2 = 0.0; u_3 = 0.0; u_4 = 0.0; u_5 = 0.0; u_6 = 0.0;
            op = 0; once = 0;
            count_ab = 0; count_anb = 0; count_nab = 0; count_nanb = 0;

            for(k = 0; k < cols; ++k) {
                if((data[i][k] == 1) && (data[j][k] == 1)) count_ab++;
                else if((data[i][k] == 1) && (data[j][k] == 0)) count_anb++;
                else if((data[i][k] == 0) && (data[j][k] == 1)) count_nab++;
                else count_nanb++;
            }
        }
    }
}

```

Figure 5-3: Main difference between the C-codes shown in the second version of the code

```

R version 2.7.2 (2008-08-25)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> dyn.load("C:/Users/admin/Desktop/Swetha files/R-C interface/Rpackage_with_full_rules_diff_writ_hardDisk$
> mychar <- c("C:/Users/admin/Desktop/Swetha files/R-C interface/Train_Data_13658.csv", "C:/Users/admin/De$
> out <- .Call("getChar", mychar)
zeros rules: 159402305
ones rules: 154144728
diff-zeros: 67956868
diff-ones: 62699291

Printed from C:
input file1
C:/Users/admin/Desktop/Swetha files/R-C interface/Train_Data_13658.csv
input file2
C:/Users/admin/Desktop/Swetha files/R-C interface/hallmarks.txt
output file1
C:/Users/admin/Desktop/Swetha files/R-C interface/Rzero.csv
output file2
C:/Users/admin/Desktop/Swetha files/R-C interface/Rone.csv

```

Figure 5-4: Output from R: Red lines are the input code and the next blue lines are the outputs after execution of the entire package after around 40 minutes

From the Figure 5-1 shown above, it can be seen that compiling the C-code was done first at the command prompt and then Figure 5-4 shows the statements to be executed to run the package. Figures 5-2 and 5-3 show parts of the C-code in the two versions of the package available.

#### **5.4. Summary**

This chapter thus shows the screenshots of the C-code. It also describes the differences in the various versions of the code. The C-code which executes the required process was thus embedded in to R to form a package which automated the process of finding good prognostic gene signatures from an entire set of genome wide interactions.

## **6 Conclusions & Prospective Work**

### **6.1. Conclusions**

Identifying the critical genes in a network would help in predicting cancer recurrence. Thus a novel network based methodology was developed which overcame the limitations of feature selection techniques. It was thus concluded that network-based techniques are capable of finding accurate and stable signatures when compared to the feature selection methods. They (network-based techniques) considered the performance of the gene interactions instead of the behavior of individual genes.

It was also seen from that implication networks are better than the currently used network-based techniques such as the correlation coefficient based and clustering based coexpression networks, Bayesian networks, and Artificial neural networks. Implication networks integrate formal logic and statistics and are thus very efficient. They also overcome the limitations of the currently used network-based techniques. Comparison of the Bayesian networks was done practically in chapter 4 from which it can be concluded that the prediction logic induced implication networks were much better in finding more gene/protein connections when compared to the Bayesian networks.

The implication network was another kind of coexpression network which was built using predication logic. Prognostic signatures were identified from the genome wide coexpression networks based on the interactions with major cancer hallmarks (E2F, EGF, EGFR, KRAS, MET, RB1, and TP53). Once the signature was obtained using genome wide implication

network, it was evaluated prognostically, clinically and structurally to make sure that the obtained signature was significant.

The Prognostic validation showed that the signature was significant with the help of KM plots, log-rank p-values, CPE values, and FDR from GSEA. The model was also compared to other classification methods from Weka. It was found that the Cox model on implication networks was much better in classifying the instances. Clinical evaluation was performed using multivariate Cox model with respect to other clinical factors. The signature was highly significant when compared to other predictors.

Structural validation was done by checking the interactions from implication networks with the interactions from Bayesian networks generated by Tetrad IV. It was seen from various web based tools that the implication networks were able to generate many more gene/protein interactions with biological relevance when compared to Bayesian networks. Weights of the implication rules were also tuned to increase the possibility of biological relevance and decrease the false discovery rates. These weights were tuned in such a way that they still include all the interactions from the curated databases. Thus all the validation methods have concluded that the signature was good. Thus the implication networks help us in finding the functional clustering between genes.

Thus it can be concluded that implication networks lead us to identify better down streamed signatures that can be used in therapeutic conditions.



## **6.2. Future Work**

A lot of prospective work can be done using this approach.

New signatures are being found using other predictive factors like Smoking status of the patients.

This network generation approach can also be tried with other cancers such as breast cancer, colon cancer, etc. Cross validation can be performed by using the signatures found in one kind of cancer to validate using datasets of other kinds of cancers. Different models can be tried with slight changes in the network generation.

In this thesis, only one set of expression values have been used to build the implication networks.

But implication networks actually have the potential to model dynamic networks with temporal relevance which can be considered in future clinical trials after surgical resections.

## List of References

1. Liu, J., and Desmarais, M. "A Method of Learning Implication Networks from Empirical Data: Algorithm and Monte-Carlo Simulation-Based Validation", IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 6, pp. 990-1004, Nov. 1997.
2. Guo, N. L., Cukic, B., and Singh, H. "Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks", ase, pp.249, 18th IEEE International Conference on Automated Software Engineering (ASE'03), 2003.
3. Sahoo, D., et al. (2008) "Boolean implication networks derived from large scale, whole genome microarray datasets". Genome Biol 9: R157.
4. Friedman, N., et al. (2000) "Using Bayesian networks to analyze expression data". *J Comp Biol* 7: 601-620.
5. Jansen, R., et al. "A Bayesian networks approach for predicting protein-protein interactions from genomic data". Science (2003) 302:449–453.
6. Aoki K., Ogata, Y., and Shibata, D. "Approaches for extracting practical information from gene co-expression networks in plant biology". Plant Cell Physiol. 2007; 48:381–390.
7. Liu, C. C., et al. "Topology-based cancer classification and related pathway mining using microarray data". Nucleic Acids Res 2006, 34(14):4069-4080.
8. Magwene, P.M., and Kim, J. (2004) "Estimating genomic coexpression networks using first-order conditional independence". Genome Biol 5: R100.
9. Goodman, P. H., and Harrell FE. "Neural networks: advantages and limitations for biostatistical modeling". In: 1998 Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association, 1999: 24-33.
10. Khoshgoftaar, T.M., Lanning, D.L., and Pandya, A.S., 1993. "A neural network modeling for detection of high-risk program". In: Proceedings of the Fourth IEEE International Symposium on Software reliability Engineering, Denver, Colorado, pp. 302-309.
11. O'Neill, M. and Song, L. (2003) "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect". BMC Bioinformatics 4:13-25.
12. Blattner, M. M., and Glinert, E. P., "Multimodal Integration", IEEE MultiMedia, v.3 n.4, p.14-24, December 1996.
13. Elo, L. L., et al. (2007). "Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process". Bioinformatics, 23, 2096-2103.
14. Hanisch, D., et al. (2002) "Co-clustering of biological networks and gene expression data". Bioinformatics 18 Suppl 1: S145-S154.

15. Choi, J. K., et al. "Differential coexpression analysis using microarray data and its application to human cancer". *Bioinformatics* 2005, 21(24):4348-55.
16. Heagerty, P.J., Lumley, T., and Pepe, M. S. (2000) "Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker *Biometrics*", **56**, 337 – 344.
17. Fawcett, T. (2006). "An introduction to ROC analysis". *Pattern Recognition Letters*, 27, 861-874.
18. Cox, D.R. "Regression models and life tables". *JR Stat Soc B*. 1972; 34:187–220.
19. Fox, J. "Cox Proportional-Hazards Regression for Survival Data". Appendix, *An R and S-PLUS Companion to Applied Regression*. 2003.
20. Shedden, K., et al. "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study". *Nature Medicine*, 2008; DOI: 10.1038/nm.1790.
21. Baudot, A., et al. (2006) "PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks". *Bioinformatics* 22: 248-250.
22. Brun, C., et al. (2003) "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network". *Genome Biol* 5: R6.
23. Kanehisa, M., et al. "KEGG for linking genomes to life and the environment". *Nucleic Acids Res*. 36, D480-D484 (2008).
24. Kanehisa, M., et al. "From genomics to chemical genomics: new developments in KEGG". *Nucleic Acids Res*. 34, D354-357 (2006).
25. Kanehisa, M. and Goto, S. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Res*. 28, 27-30 (2000).
26. Kanehisa, M. "Toward pathway engineering: a new database of genetic and molecular pathways". *Science & Technology Japan*, No. 59, pp. 34-38 (1996).
27. Ulitsky, I. and Shamir, R. 2007. "Identification of functional modules using network topology and high-throughput data". *BMC Systems Biology*, 1:8, 2007.
28. Paik, S., et al. "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer". *New England Journal of Medicine*. 2004: 351:2817-2826.
29. Gonen, M., and Heller, G. (2005) "Concordance probability and discriminatory power in proportional hazards regression". *Biometrika* 92, 965-970.
30. Subramanian, A., et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *PNAS*. 102, pg 15545-15550.
31. Ping, H. E., et al. "Multivariate analysis by Cox Proportional Hazards Model on prognoses of patients with bile duct carcinoma after resection". *CMJ* 2002; 115(10): 1538-1541.

32. Alberg, A. J., et al. 2004. "The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests". *J. Gen. Intern. Med.* 19:463-468.
33. Ideker, T., and Sharan, R. (2008). "Protein networks in disease". *Genome Res.* 18,644–652.
34. Karlebach, G., and Shamir, R. "Modelling and analysis of gene regulatory networks". *Nat.Rev.Mol.Cell Biol.* 2008 Oct; 9(10):770-80.
35. Gordon, G. J., et al. "Validation of genomics-based prognostic tests in malignant pleural mesothelioma". *Clin Cancer Res.* 2005 Jun 15;11(12):4406-14.
36. Jensen, L. J., et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms". *Nucleic Acids Res* **37** (Database issue): D412-6.
37. Von Mering, C., et al. (2007). "STRING 7--recent developments in the integration and prediction of protein interactions". *Nucleic Acids Res* **35** (Database issue): D358-62.
38. Von Mering, C., et al. (2005). "STRING: known and predicted protein-protein associations integrated and transferred across organisms". *Nucleic Acids Res* **33** (Database issue): D433-7.
39. Von Mering, C., et al. (2003). "STRING: a database of predicted functional associations between proteins". *Nucleic Acids Res* **31** (1): 258-261.
40. Snel, B., et al. (2000). "STRING: a web-server to retrieve and display the repeatedly occurring neighborhood of a gene".
41. Tornow, S., and Mewes, H. W., "Functional Modules by Relating Protein Interaction Networks and Gene Expression", *Nucleic Acid Research*, vol. 31, no. 21, pp. 6283-6289, 2003.
42. Larsen, P., et al. "Correlated discretized expression score: a method for identifying gene interaction networks from time course microarray expression data", *Proceedings of the 28th International Conference of IEEE Engineering in Medicine and Biology Society (EMBS)* (2006). pp.5842-5845.
43. Zhang, B., Horvath, S. (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", *Statistical Applications in Genetics and Molecular Biology: Vol. 4 : Iss. 1, Article 17.*
44. Boger, Z., "Artificial neural networks methods for identification of the most relevant genes from gene expression array data", *Proceedings of the International Joint Conference on Neural Networks*, (2003). vol.4, pp. 3095- 3100.
45. Xu, Y., et al. "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer". *Cancer Res* 2002;62 (12): 3493–7.
46. Keedwell, E., and Narayanan, A. "Discovering Gene Networks with a Neural-Genetic Hybrid", presented at *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2005, pp.231-242.
47. Kim, Y., Street, W. N., Menczer, F. (2002). *Feature selection in data mining* (pp. 80-105). Hershey, PA: Idea Group Publishing. (Book Chapter).

48. Kaplan, E.L. & Meier, P. (1958). "Nonparametric estimation from incomplete observations". *Journal of the American Statistical Association* 53: 457–481.
49. Thalamuthu, A. et al. (2006) "Evaluation and comparison of gene clustering methods in microarray analysis", *Bioinformatics*, 22, 2405-2412
50. Kevin Gurney. "An Introduction to Neural Networks". UCLPress, 1997.
51. Potti, A., et al. "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer". *N Engl J Med*. 2006;355:570-580.
52. Chen, H-y., et al. "A five-gene signature and clinical outcome in non-small cell lung cancer". *N Engl J Med* 356: 11-20. 2007.
53. Beer, D.G., et al. (2002). "Gene-expression profiles predict survival of patients with lung adenocarcinoma". *Nat. Med.* 8:816-824.
54. Boutros, P. C., et al. (2009)." Prognostic gene signatures for non-small-cell lung cancer". *Proc. Natl. Acad. Sci. USA* 106: 2824-2828.
55. Bhattacharjee, A., et al," Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses". *Proc. Natl. Acad. Sci. USA* **98** (2001), pp. 13790–13795.
56. Raponi, M., et al. (2006) "Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung". *Cancer Res* 66:7466–7472.
57. Lau, S. K., et al. (2007) "Three-gene prognostic classifier for early-stage non small-cell lung cancer". *J Clin Oncol* 25:5562–5569.
58. Lu, Y., et al. (2006) "A gene expression signature predicts survival of patients with stage I non-small cell lung cancer". *PLoS Med* 3:e467.
59. Guo, N. L., et al. "Constructing Molecular Classifiers for the Accurate Prognosis of Lung Adenocarcinoma". *Clin. Cancer Res.* (2006) 12:3344-3354.
60. Bild, A. H., et al. (2006) "Oncogenic pathway signatures in human cancers as a guide to targeted therapies". *Nature* 439:353–357.