Significant gene array analysis and cluster-based modeling for disease class prediction

Myrine Barreiro-Arevalo
Hansapani Rodrigo, Ph.D.

University of Texas Rio Grande Valley

October 2020

Introduction

- Gene expression analysis has been of major interest to biostatisticians for many decades. Such studies are necessary for the understanding of disease risk assessment and prediction, so that medical professionals and scientists alike may learn how to better create treatment plans to lessen symptoms and perhaps even find cures.
- Identification of genes related to diseases and other biological problems has been a concern given phenotypic co-expression and pathway development in a genome.

Previous Studies

- Several techniques have been created to analyze gene expression data with the introduction of microarray technology.
- One popular one being random forest modeling, a popular supervised machine learning tool that has recently gained traction in the field of computational biology due to its ability to analyze DNA microarrays and handle thousands of data points simultaneously.

- Unfortunately, random forest modeling alone will only analyze genes one-by-one.
- It is generally well known that genes do not act by themselves, but rather in groups of genes, or “clusters”. These clusters are usually closely linked sets of genes that are all necessary to perform a function or express a phenotype.

Objectives

Find clusters of co-expressed genes and identify the effect of clustering classification based on previous knowledge in gene expression data/microarray data.

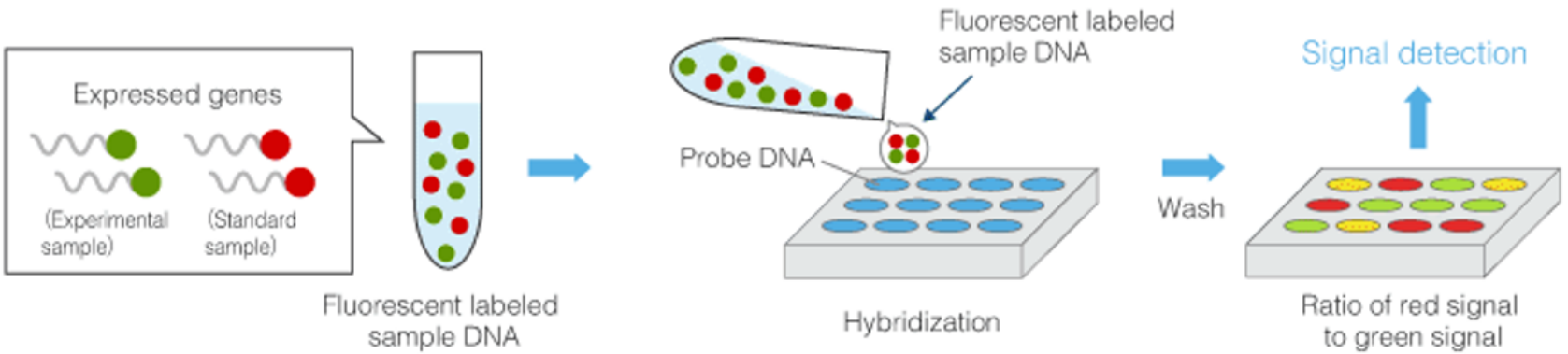
Investigate various gene expression analyses and machine learning techniques for disease class prediction, as well as assess predictive validity of these models and uncover differentially expressed (DE) genes for their relevant datasets.

Our models to be addressed are:

1. Simple Random Forest
2. Gene eXpression Network Analysis (GXNA),
3. Significant Analysis of Microarrays (SAM) is used to identify potential disease biomarkers

Microarray Technology

- DNA microarray technology has had a specific significant breakthrough in the field of molecular biology because of its capability of handling thousands of gene expression data simultaneously.
- Our datasets are using the Affymetrix Human Genome U133A platform (GPL96).



Summary of the Subjects

Multiple gene expression datasets will be used to test model accuracy and will be obtained from public access Gene Expression Omnibus (GEO).

Dataset 1 230 subjects: 144 lymph-node negative relapse free patients and 86 lymph-node negative patients that developed a distant metastasis (breast cancer)

Dataset 2 197 subjects: 90 smokers with abnormalities found in the bronchial epithelium (lung cancer) and 97 smokes without abnotmalities found

Dataset 3 39 subjects: 24 post-mortem brain tissue samples from individuals diagnosed with Parkinson's disease and 15 control individuals

Initial Experimental Procedures

- Typically, experiments compare two or more phenotypes or disease states with many subject replicates. Each subject replicate measures expression data for a large number of genes.
- Standard analyses first start with filtering and normalizing gene data, with computations following for each gene to compare the expression levels between the different phenotypes or disease states.
- Finally the genes are sorted in increasing order and the most significant ones are used for experimental validation.

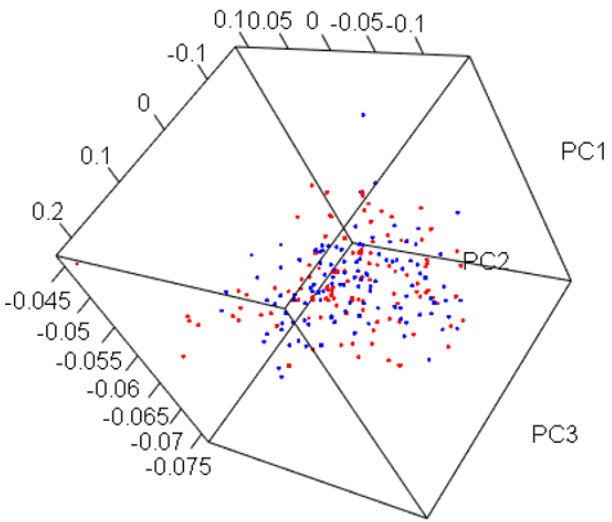
Non-Specific Filtering of Genes

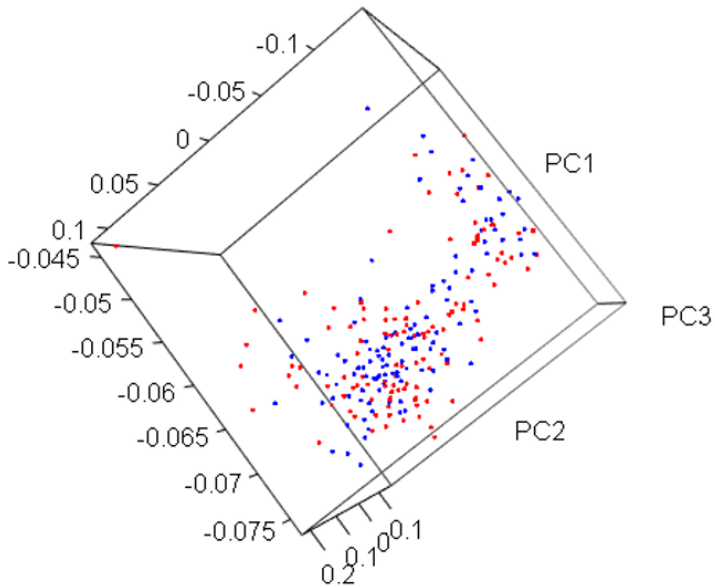
- Gene expression data that did not have enough variance to be expressed or informative in classification are filtered out using GCRMA, which adjusts for background noise and non-specific binding.
- At least 20% of the sample genes with unlogged normalized intensity greater than 100.
- coefficient of variation (sd/mean) is between 0.7 and 10.
- After the GCRMA normalization, we get log2 gene expression data.

Principal Component Analysis

- Filtered log2 gene expression data was used to fit a linear model with weighted least squares with empirical Bayes moderation of the standard error. This approach is well suited to identify differentially expressed genes with are not normally distributed when the expression values differ between genes.
- We apply the Benjamini-Hochberg correction for multiple comparison testing.
- The genes with adj. p value less than 0.01 were identified as differentially expressed.
- PCA was preformed to identify subjects with similar gene profiles.

Principal Component Analysis





Significant Analysis of Microarrays

- SAM identifies statistically significant genes by gene specific t-tests and computes a new statistic, d_j for each gene j, which measures the strength of the relationships between gene expression and the response variable (disease state).
- Use of permutations accounts for correlation of genes and avoids parametric assumptions like normality.
- A disease state response uses Two classes unpaired, where the measurement units of the classes are different i.e. control and treatment patient groups.

Significant Analysis of Microarrays

Dataset 1

Up-regulated Genes			
Gene Name	Gene ID	Score	Fold Change
TRAF5	7188	2.26789	2.26669

Dataset 2

Up-regulated Genes			
Gene Name	Gene ID	Score	Fold Change
PITPNA	5306	1.74898	1.31351
MFN2	9927	1.72217	1.33682
SRPRA	6734	1.71707	1.31622
FAM120A	23196	1.67434	1.30648
CIRBP	1153	1.49008	1.38110

Down-regulated Genes			
Gene Name	Gene ID	Score	Fold Change
TMED2	10959	-1.49623	0.74534
HSBP1	3315	-1.46084	0.71806

Dataset 3

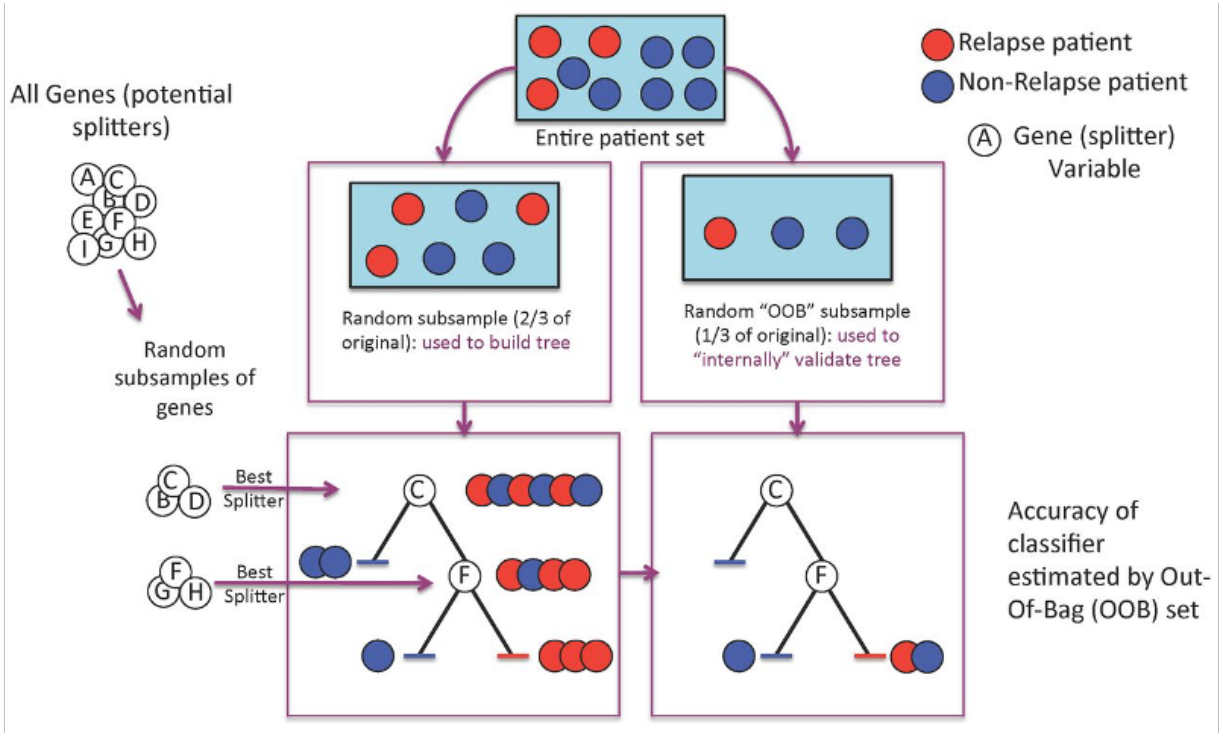
Up-regulated Genes			
Gene Name	Gene ID	Score	Fold Change
TARDBP	23435	3.70084	3.38299
CCL5	6352	3.69190	4.35597
ATF4	468	3.62920	4.75927
CCND2	894	3.62010	3.78969
HSPD1	3329	3.61627	3.08266

Down-regulated Genes			
Gene Name	Gene ID	Score	Fold Change
C11orf58	10944	-4.35889	0.27277
ACTR2	10097	-4.05302	0.35632
DNAJB1	3337	-4.04651	0.28559
MARCKSL1	65108	-3.89071	0.38643
GABARAP	11337	-3.59386	0.33807

A Random Forest Classification Model

- A random forest is a collection of decision trees.
- A decision tree builds classification model in the form of a tree structure.
- It breaks down a dataset into homogeneous subsets while incrementally developing a decision tree.
- Each decision tree in a RF model:
  - is built using a bootstrap sample of observations.
  - a best split is chosen from a random subset of predictors rather than all of them.
- Prediction accuracy is measured using Out-Of-Bag (OOB) data.

Random Forest Classification Model



Random Forest Classification Model

- Random forest is a classification method well suited for microarray data with a good predictive performance.
- Identification of the most substantial set of genes which depict significant variation between the disease state subjects.
- Shows excellent performance even when most predictive variables are noise.
- Suits well for small "n" and large "p" problems and when there are more than two classes.
- Returns measures of variable importance.

Gene eXpression Network Analysis

- Substantial gene identification by one-at-a-time models like random forest is not the most appropriate technique as many genes are correlated and function together.
- GXNA computes a score that measures to what extent a gene or set of genes is differentially expressed.

Gene eXpression Network Analysis

- Substantial gene identification by one-at-a-time models like random forest is not the most appropriate technique as many genes are correlated and function together.
- GXNA computes a score that measures to what extent a gene or set of genes is deferentially expressed.

Gene eXpression Network Analysis

- A random forest model is created with these highly informative clusters from the GXNA.
- This provides relatively efficient biomarker identification.

Model Details

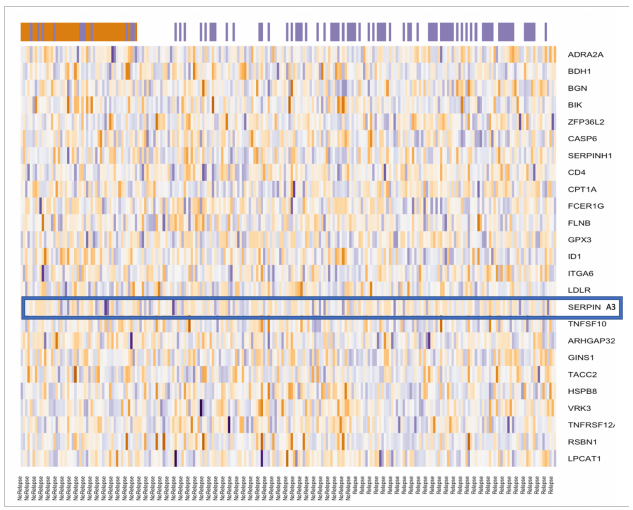
Accuracy			
Method	Dataset 1	Dataset 2	Dataset 3
Simple RF	0.54	0.67	0.82
RF w/ GXNA	0.58	0.71	0.87

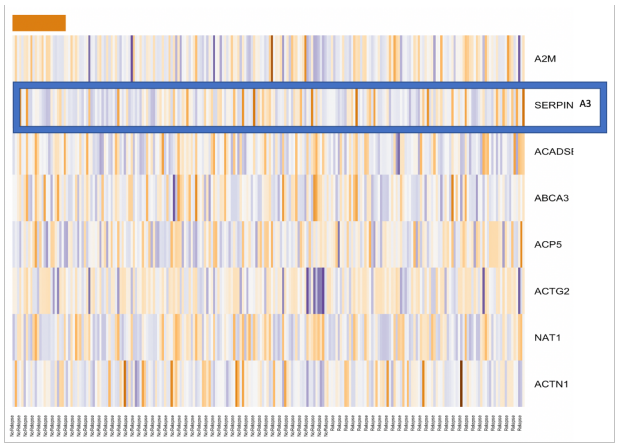
Sensitivity			
Method	Dataset 1	Dataset 2	Dataset 3
Simple RF	23.26	71.11	60.74
RF w/ GXNA	34.88	81.11	83.28

Specificity			
Method	Dataset 1	Dataset 2	Dataset 3
Simple RF	72.91	60.3448	58.99
RF w/ GXNA	67.36	55.17	79.17

Heatmaps

Dataset 1

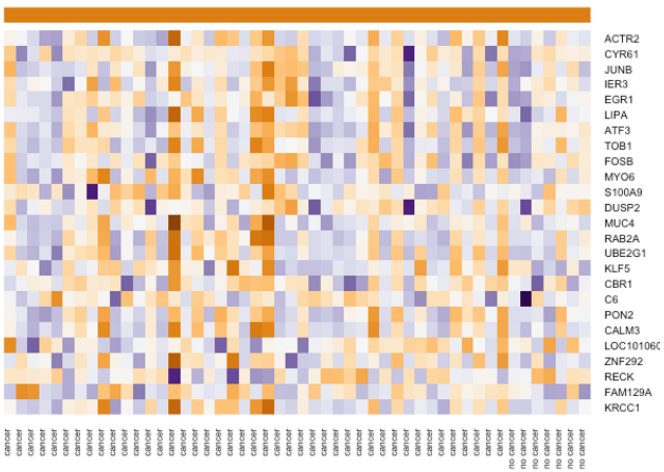


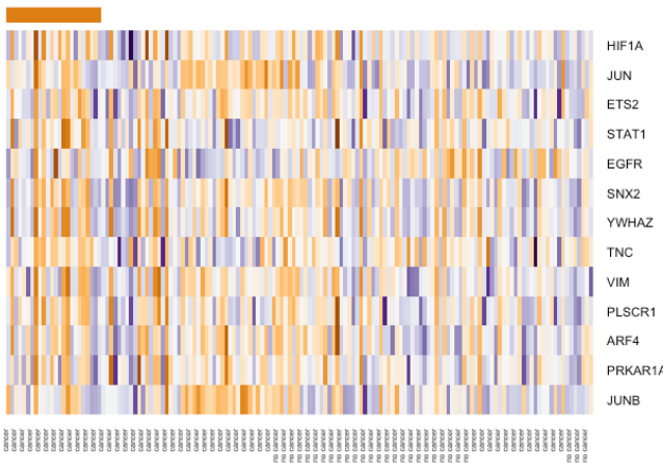


- SERPINA3 gene was present and down-regulated in relapse patients in both the Random Forest and Random Forest with Clusters model, showing that it could be a potential biomarker for breast cancer relapse.

Heatmaps

Dataset 2





- JUNB gene was present and down-regulated in cancer patients in both the Random Forest and Random Forest with Clusters model, showing that it could be a potential biomarker for lung cancer.

Final Thoughts

Overall, the model with the highest accuracy and sensitivity was determined to be the Random Forest model with highly ranked clusters. This shows that it has good potential to be used as a biomarker identification model and is a possible approach to disease classification.

Future Analysis

Our next step in the analysis is:

- LASSO regression
- Bayesian Neural Network

References

- J.K. Lee, P.D. Williams, and S. Cheon, “Data Mining in Genomics", Clin Lab Med. **28**, 145-viii (2008)
- H. Pang, A. Lin, M. Holford, B.E. Enerson, B. Lu, M.P. Lawton, E. Floyd, and H. Zhao, “Pathway Analysis using Random Forests Classification and Regression", Bioinformatics **22**, 2028-36 (2006)

Thank You!