Non-linear Reconstruction of Genetic Networks Implicated in AML Pathology

CS 229 – Autumn 2014

Aaron Goebel *goaaron*
Mihir Mongia *mmongia*

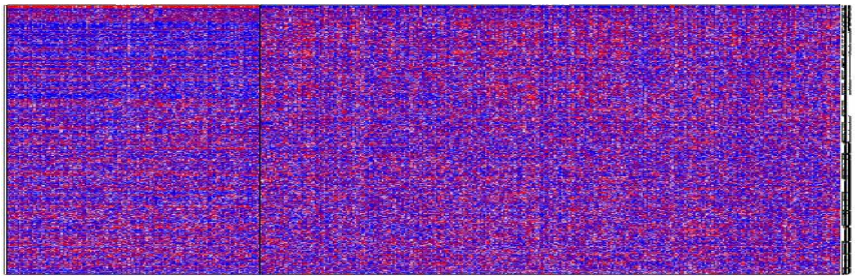
1 Problem Introduction

The origins and underlying mechanisms of cancer are widely unknown. It has become clear that to truly understand and manipulate cancer, one must understand how genes interact with each other and the environment. The mechanisms behind protein creation or the creation of any biological matter are extremely complicated. Sometimes a gene may encode a protein that activates another gene! Physically and dynamically monitoring all these interactions is impossible at the moment. Thus data analysts often create models of gene regulatory networks given gene expression data in order to point experimentalists in the right direction. Often the data sets that are used are very "skinny" in the sense that there will be data for 500 patients and for each patient there will be data on the order of 10,000 genes. Clearly this is not a lot of data for each gene especially considering that gene expression data can be quite noisy. A data analyst must address these challenges.

In particular, we focus on a dataset of gene expression data in Acute Myeloid Leukemia (AML). Furthermore, we focus on the presence or absence of 2 certain proteins NPM1 and FLT3-ITD, which are good prognosticators for patient outcome. We would be further along in understanding the origins of certain cancers if we could somehow pinpoint which genes or which gene interactions cause the existence of the NPM1 and FLT3_ITD proteins.

2 Dataset

We were given a dataset from Andrew Gentles, a bioinformatics researcher at the Stanford University lab for integrative cancer biology. This data is gene expression data for Acute Myeloid Leukemia or AML. The data is derived from a gene expression microarray profiling 524 cases of de novo AML across expression levels with 17,788 genes (comparisons of cases with double and single CEBPA mutations versus those with wild type CEBPA). The data set also comes with binary classification entries for the presence of NMP1 and FLT3-ITD expression in each observations.



Microarray heat-map for gene expression. (Left) NMP1 Positive observations, (Right) NMP1 Negative expression observations.

3 Features and Processing

We initially could not discriminate among the 17,000 genes in terms of what would be better predictors for the expression of NMP1 or FLT3 without any domain knowledge. In terms of traditional data processing, we reduced the dimension of this data set 25 by using PCA. This did not prove to be very

good for classification purposes. We also tried Lasso methods which will be explained in the following section. The most important pre-processing we did is that we calculated the mutual information between genes and created an adjacency matrix out of this. This is done via ARACNE, which is a software developed by computational biologists at Columbia University. The idea is essentially that if two genes have high mutual information, then they are likely to be connected (if we were to describe the interactions of genes with a graph). Mutual information is used as a well-behaved indicator of non-linear dependencies that cannot be caught by other metrics such as the Pearson correlation, Spearman-ρ, or Kendall-τ (it is possible **–and often likely–** that two genes have an interesting relationship and yet have a correlation of zero). We also investigated implementation of the more novel and dynamic metric of distance correlation, however that endeavor proved too computationally expensive and drawn out to complete within the quarter.

4 Models and Techniques

PCA – Since we have nearly 17,000 genes in our data set, we try to see if can reduce our data to a smaller, more tractable subset. This method by nature eliminates the effect of one gene on a binary output (presence of NPM1 or FLT3 Protein). However, it is still a good initial tool to see if one can predict a binary output when the dimension of the data is smaller than the number of patients (i.e. avoiding this issue of a "skinny" dataset). We reduce our 17, 000 dimension data set to a 25-dimensional space.

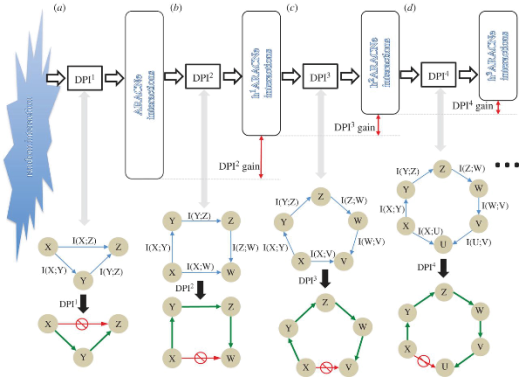
Logistic Regression – We use this modality as a primary classifier algorithm.

Lasso – We implemented the L_1 penalized regression method to generate a sparse regression of only the most significant genes in our feature space so as to generate a more interpretable model that does not suffer with over-fitting due to the large excess of explanatory variables in relation to the number of observations. Implementations of the combined L_2 elastic-net were also tested, but they proved unstable for dimensionality reduction and did not produce large increases in classification success.

Enriched Lasso – The enriched Lasso extension is an attempt to bypass some limitations of the Lasso, namely that the algorithm is prone to random exclusion of one of two similarly valued genes. It is somewhat hard to verify the statistical significance of the parameters chosen by Lasso. The enriched method works by finding t-statistics of genes differentially expressed about classes, and then correcting these for false discovery rate (FDR). The resulting "q-value" is then utilized as an additional weighting parameter for individual genes in the Lasso. We use the heuristic of applying a weight of $(-\log q_i)$ to the i_{th} respective gene in the Lasso formulation.

$$\hat{\boldsymbol{\beta}}(\text{AdaEnet}) = \left(1 + \frac{\lambda_2}{n}\right)\left\{\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1^* \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|\right\},$$

ARACNE Model – We utilize the ARACNE algorithm to calculate pair-wise mutual information of genes. The algorithm imparts further insight into our network by applying the Data Processing Inequality (DPI) on the network graph to infer the most likely path of information flow by removing false candidate interactions, i.e., triplet cycles are severed by ranking to elucidate more meaningful topological structure. The figure to the right displays the DPI procedure.



Calculating Total Number of Paths – Given an adjacency matrix A in which the i,j$_{th}$ entry corresponds to the weighted connection between gene i and gene j we can find the number of 2-step path lengths between node i and node j by simply calculating A^2. In particular:

$$A = \begin{matrix} 1 & 0 & .5 \\ 0 & 0 & .8 \\ .5 & .8 & 0 \end{matrix}$$

We can see that there is one 2 step path length from node 2 to node 1. First one goes to from node 2 to node 3. Then one can go from node 3 to node 1. Fortunately, we can measure this by simply looking at the (2,1) entry of A^2. The value here will be non-zero and effected by the strength of the connections between each path step. For example if we replaced each non-zero entry here with a one, the (2,1) entry of A^2 would be 1. However in our case the (2,1) entry of A^2 will be four-tenths.

This matrix A^2 is simply the number of 2-step path lengths. What about the number of 3 –step path lengths? Or 10? Does a 10-step path length really matter? It would make sense that the greater number of steps the less we care about the connection. Thus we make our own metric Atotal , and Atotal is defined as

$$Atotal = A + A^2 * \frac{1}{2} + A^3 * \frac{1}{6} + A^4 * \frac{1}{24}$$

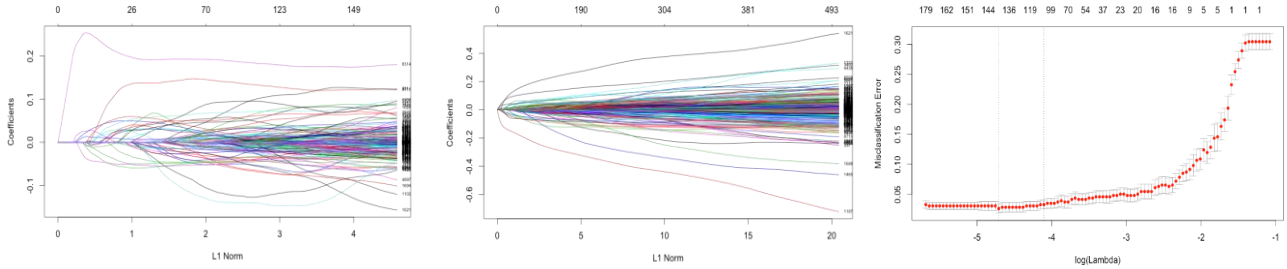
We then examine the nodes with the most connections by summing up the entries in the column corresponding to a certain node.

Eigenvector Centrality – We also use an algorithm similar to page rank. Given an adjacency matrix A, we can find the most connected nodes by finding the eigenvector of matrix A corresponding to the largest eigenvalue in A. By the *Perron-Frobenius* theorem, this eigenvector has only non-zero entries and each entry in the vector corresponds to the entry's importance. It turns out that the method mentioned before and this method give us very similar results and so we will only show the plots for one method in the Results section.

5 Results and Discussion

PCA: We reduced the size of the dataset to dimension 25 using traditional PCA. When we trained a logistic regression classifier on this data we could only predict with accuracy of 85%.

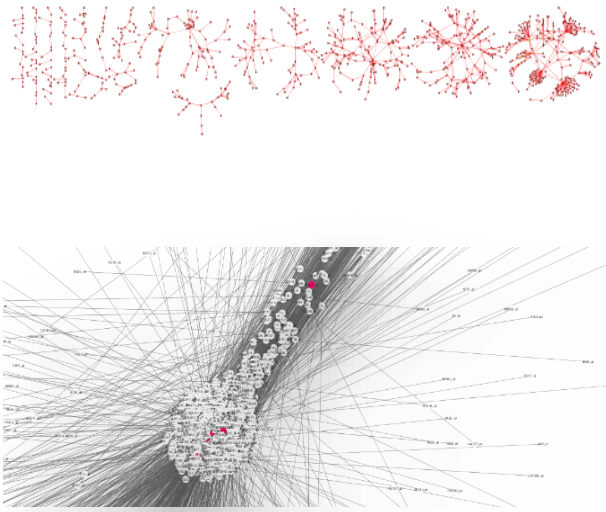
LASSO: Given the Lasso data, we found that there was one gene whose coefficient was always significantly higher than the others. We trained logistic regression just on the "Hypothetical LOC404266" gene and were able to get a testing error of 15%. The unmodified LASSO reduced our feature space to 103 significant genes with 6% classification error while the enriched LASSO reduces the feature space to a mere 13 genes, but at the cost of a 38% classification error.



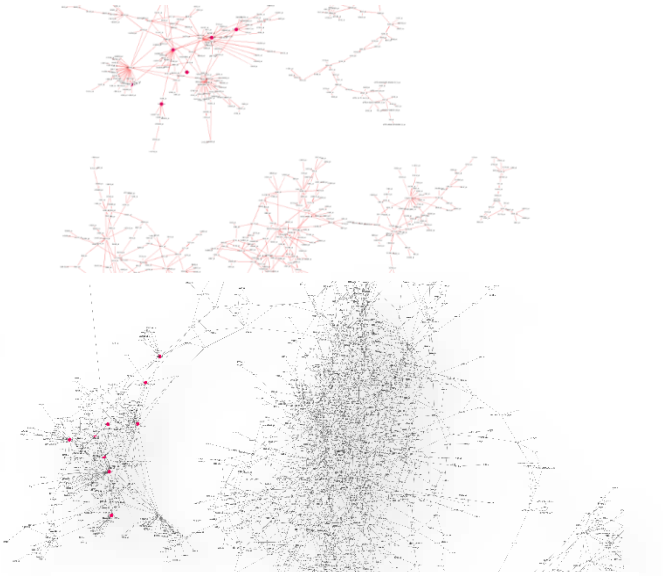
(Left) Lasso path iteration, (Middle) FDR enriched Lasso path
(Right) Misclassification cross validation error of the unenriched Lasso with optimal lambdas.

Learning Algorithm	Training Error	Testing Error
Principal Components Analysis (25 components)	0% (100 Samples)	14% (300 samples)
Lasso $\lambda_2 = 0$ **(103 Genes)**	0% (100 Samples)	6.5% (200 samples)
Enriched Lasso $\lambda_2 = 0$, $\omega_1 = FDR\ corrected\ p - values$ from Wilcox-test **(12 Genes, no LOC404)**	22%(100 Samples)	38.6% (300 samples)
Logistic Regression with only hypotehetical_LOC404266 **(1 Gene)**	8% (100 Samples)	15% (300 samples)

Network Graphs:

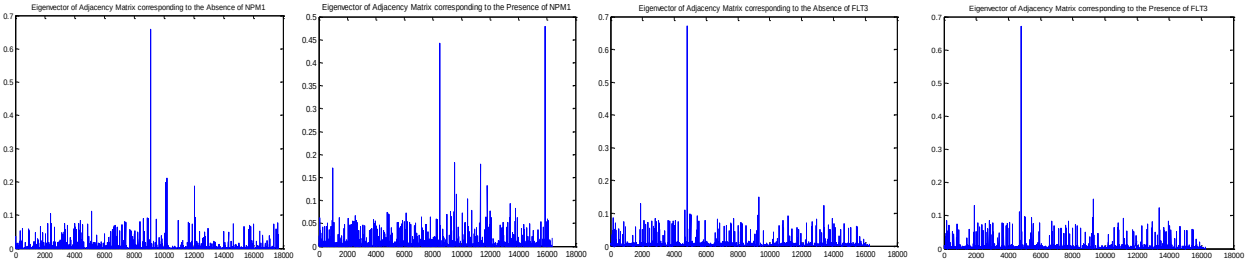


Top: The MI weighted force-directed network for the NMP1 negative class. *Bottom:* Force-directed network for the NMP1 positive class. Genes of high-implication are circled.



Top: Organic network (plain visualization) for the NMP1 negative class. *Bottom:* Organic network for the NMP1 positive class. Genes of high-implication are circled.

Visually, there are obvious differences between the networks of a NMP1 expresser and non-expresser. Namely, the expresser exhibits signs or rampant connectivity with the principal genes at the center of the force-directed network and as hubs of the organic network. The non-expresser displays stark delineations between sub-networks—likely a sign of natural segregation rather than the unnaturally high amount of cross-talk in our expresser network.



We did an Eigenvector Centrality analysis (Above figure) on the graphs generated by ARACNE. The results were interesting. Although the FLT3 eigenvectors did not change much when FLT3 was absent and when FLT3 was present, in the case NPM1, we found some interesting variations. One can see in the data above that in the presence of NPM1 a few genes became more connected than they were in the absence NPM1. In the case of NPM1, the genes that became more connected were CCDC64, CYP2W1 , CCL27.

6 Conclusion

We initially started out with low-hanging fruit by applying linear techniques. One particularly fruitful method we tried was LASSO. We found that one gene was highly predictive of the presence of the NMP1 protein. In fact, if we constructed a logistic regression model using simply that one gene, "Hypothetical LOC404266". We were able to have an 85% percent prediction rate. In addition, we constructed a theoretical graph corresponding to connections of genes. We did this by measuring mutual information in gene expression data. We constructed graphs corresponding to the presence of NMP1 and graphs corresponding to the absence of NMP1. We then used an algorithm similar to PageRank to find the most highly connected genes. We found that a few genes became suddenly highly connected in the presence of NMP1. These genes are CCDC64, CYP2W1 , CCL27. These genes along with "Hypothetical LOC404266" are genes we would recommend to explore experimentally. Of course, the former of these suggestions is only a good recommendation if the graph based on the mutual information graph is relatively close to accurate. On synthetic graphs the ARACNE algorithms has proved to be quite successful. Thus at this point we are relatively optimistic on our recommendations.

Future

If there was more time to pursue this project more seriously, we would look into differences between the genetic networks such as unique gene paths. In this project we simply studied how the central nodes changed. Secondly we would acquire more data on different types of cancer to better understand the differences between these graphs. We would also try a more robust metric than mutual information, namely distance correlation (but that was going to take us 6000 hours on the campus servers). Finding a metric that somehow is more suited for biological relationships would be ideal. For example maybe it's the case the in biology there is a common type of relationship between genes that are interacting. If we could create a metric suited for that, then we could do a much better job in find genes that are connected.

7 References

[1] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, A. Califano Reverse engineering cellular networks. Nature Protocols, 1 (2) (2006), pp. 662–671

[2] J. Huang, S. Ma, and C.-H. Zhang, *Adaptive Lasso for sparse high-dimensional regression models*, Statist. Sinica. 18 (2008), pp. 1603–1618.

[3] Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proceedings of the National Academy of Sciences 98(9): 5116–5121.

[4] N. Rapin, B. Porse, et al. (2013) *Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients, Blood: 123 (6)*

[5] YU Hui, MITRA Ramkrishna, YANG Jing, LI YuanYuan, ZHAO ZhongMing. Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. SCIENCE CHINA Life Sciences, 2014, 57(11): 1090-1102.