

# The Induction and Analysis of Gene Networks

Gary Livingston

University of Massachusetts—Lowell  
One University Avenue  
Lowell, Massachusetts 01854  
978-934-4694

gary@cs.uml.edu

Guangyi Li

University of Massachusetts—Lowell  
One University Avenue  
Lowell, Massachusetts 01854  
978-934-4694

gli@cs.uml.edu

Liwu Hao

University of Massachusetts—Lowell  
One University Avenue  
Lowell, Massachusetts 01854  
978-934-4694

lhao@cs.uml.edu

Xiao Li

University of Massachusetts—Lowell  
One University Avenue  
Lowell, Massachusetts 01854  
978-934-4694

xili@cs.uml.edu

## ABSTRACT

We report on the use of programs based upon rule induction for the induction of gene influence networks from microarray data and for analyzing the resulting networks. We use these programs to induce gene networks from a dataset created by merging normalized versions of the Beer et al. and Battacharjee et al. datasets made available for the 2003 CAMDA conference. We verified our findings by comparing them to lists of known oncogenes. Our results suggest that (1) normalization can be used to combine microarray datasets to yield a dataset with a larger number of cases, which facilitates discovery from the cases, and (2) our programs can be used to induce gene influence networks from microarray datasets and to use the induced networks to identify statistically significant differences in subgroups of the datasets. For example, our programs have rediscovered many known oncogenes from the combined dataset.

## General Terms

Algorithms, experimentation, evaluation

## Keywords

Bioinformatics, microarray data analysis, pathway induction, pathway analysis

## 1. INTRODUCTION

The induction and analysis of gene influence networks generated from microarray gene expression data may greatly aid the understanding and cure of gene-based diseases such as cancer. We have been studying the problem of using microarray gene expression data to create gene influence networks. We have devised a novel set of programs for (1) inducing gene networks from gene expression data that are based upon rule induction and (2) analyzing the induced networks. The first program, InduceNet, induces networks from gene expression data; the second program, CompareNet, performs an edgewise comparison of two networks; and a third program, ComparePop, uses a given network to compare the “fit” of the network to two populations. In the study reported here, InduceNet and ComparePop were used to analyze the CAMDA datasets after the datasets were normalized and merged.

With the CAMDA datasets, as is the case with the state of microarray databases, we have the problem and opportunity of having four datasets that we may use. The problem is how to leverage the additional cases that four datasets provide beyond merely using each dataset to verify the results of the others, which in itself is difficult, because the genes measured in each dataset may differ widely, the samples come from different populations, and the method used to obtain the expression levels varies, so the expression levels also vary. The opportunity if we could somehow combine the data sets, we have more cases available for learning!

Our approach was to normalize the expression levels for each gene in each dataset using the mean and standard deviation of the same “reference” subpopulation from each dataset. We use cases with stage 1 tumors as our reference subpopulations for the datasets because there are a reasonable number of these cases in each dataset. We used this approach to combine the Beer, D. et al. [1]

and Battacharjee, A. et al. [2]. We did not use the third data set from Wigle, D. et al. [7] because of its high degree of missing information and small size, and we did not use the fourth data set, from Garber, M. E. et al. [3] because we could not convert the gene descriptions into gene names which we needed for our approach to combining the data sets.

## 2. Methods

### 2.1 Data pre-processing

Our data processing involved the following steps:

1. Identify and select the genes that were measured in the three databases we used. All other genes were omitted.

2. Normalize the values for these genes:

For each selected gene,

For each database,

Calculate the mean and standard deviation of the gene's values for cases with stage 1 tumors

Normalize the values for that gene using mean and standard deviation calculated above

3. Merge the Beer et al. and Battacharjee et al. datasets
4. Select the 500 genes in the combined datasets that varied the most. We calculated the variance of each gene in the combined data, using their normalized values to calculate the variance, and selected the 500 genes with the most variance

### 2.2 InduceNet: inducing gene networks

Our program for inducing gene networks from microarray data, InduceNet, uses a discovery program called HAMB [5] to generate rules predicting the expression levels of each gene by using the expression levels of the other genes. Rule sets using one gene to predict another with an accuracy (proportion of correct predictions) exceeding a given threshold form the edges of the induced network.<sup>1</sup>

HAMB [5] is a supervisor program to a rule induction program called RL [7]. HAMB has many features which make it desirable for analyzing gene expression data: (1) the user may specify multiple "target" attributes for which rules sets will be induced, (2) HAMB automatically selects the feature set and parameters for each of the target attributes using search and heuristics, and (3) HAMB does some post processing of the induced rules, such as pruning and grouping similar rules. Moreover, when it is available, HAMB can use domain knowledge to improve the quality of its reported rules and rule sets.

One of HAMB's methods for post processing rules is to group them into *rule families*. These are groups of rules where changing

the value of one attribute on a rule's left-hand side yields in a consistent change in the value being predicted. The consistency of the rules in the family increases confidence in the rules. Table 1 presents some of the stronger rule families HAMB discovered from the Beer et al. dataset for tumor differentiation, P53 nuclear accumulation, and tumor stage. Rules 1–3, 4 and 5, 6 and 7, 8–10, and 11–13 form 4 rule families. The p-values for all rules in Table 1 are less than 0.00001. When we validated rules 11–13 using the Battacharjee et al. data, the accuracy of the three rules used together to predict tumor stage was 59%, with 60% coverage. The p-values of these rules on the Battacharjee et al. data were <0.0001, 0.119, and 0.0115, respectively, which suggests that the trend exists in the Battacharjee et al. data. Used together, rules 1–3 use CDKN3 to predict tumor differentiation with 60% accuracy and 60% coverage, rules 4 and 5 use KLF5 to predict P53 nuclear accumulation with 72% accuracy and 41% coverage, rules 6 and 7 use PSG5 to predict tumor stage with 66% accuracy and 40% coverage, rules 8–10 use KRT6A to predict tumor stage with 59% accuracy and 35% coverage, and rules 11–13 use CP to predict tumor stage with 59% accuracy and 60% coverage.

HAMB's ability to automatically set up and run rule induction tasks and to group its results into rule families was critical to the induction of gene networks. HAMB's ability to automatically perform rule induction tasks and organize the results allowed it to set up and run the 500+ rule induction runs needed for the 500 genes plus patient attributes. Manually setting up 500+ induction runs would have taken weeks!

**Table 2. Summary of rule families discovered by HAMB. Please refer to the text for an explanation of the columns.**

ID	RULE	TP	F P	SENS	PP V
1	(CDKN3 MIN) ==> (DIFFERENTIATION WELL)	20	18	0.43	0.5 3
2	(CDKN3 MED) ==> (DIFFERENTIATION MODERATE)	30	8	0.37	0.7 9
3	(CDKN3 MAX) ==> (DIFFERENTIATION POOR)	20	20	0.48	0.5 0
4	(KLF5 MAX) ==> (P53_NUCL_ACCUM -)	40	0	0.29	1.0 0
5	(KLF5 MIN) ==> (P53_NUCL_ACCUM +)	16	22	0.50	0.4 2
6	(PSG5 MIN) ==> (STAGE 0)	12	24	0.60	0.3 3
7	(PSG5 MAX) ==> (STAGE 1)	38	2	0.28	0.9 5
8	(KRT6A MAX) ==> (STAGE 3)	18	22	0.47	0.4 5
9	(KRT6A MIN) ==> (STAGE 0)	12	26	0.60	0.3 2
10	(KRT6A MED) ==> (STAGE 1)	38	0	0.28	1.0 0

<sup>1</sup> Both HAMB and RL are publicly available. Public release versions of InduceNet, CompareNet, and ComparePop will be available soon.

11	(CP MIN) ==> (STAGE 0)	12	26	0.60	0.3 2
12	(CP MAX) ==> (STAGE 3)	20	20	0.53	0.5 0
13	(CP HIGH) ==> (STAGE 1)	36	2	0.27	0.9 5

### 2.3 CompareNet: comparing networks

CompareNet creates a “difference network” from two networks by comparing the edges in the two networks, thus creating a new network comprised of the edges that differ between the networks as follows: edges that appear in the first network but not in the second, edges that appear in the second network but not in the first, and edges that appear in both networks, but with different correlations (e.g., a positive correlation in the first network and a negative correlation in the second).

### 2.4 ComparePop: comparing populations

We realized that merely inducing a gene network from microarray datasets involving cases with cancer and without would not identify the gene interactions that are particular to cancer. Nor would gene networks induced from only cases involving cancer: while gene networks induced from cancer cases should indicate interactions among the genes that are related to the cancer, there would still be many interactions that would be discovered that would be related to normal gene interactions, and distinguishing among them would be difficult. Therefore, we designed ComparePop, which takes a given network and two populations and compares the “fit” of each edge of the network to each population, then performs a statistical test to determine if the fit differs significantly between the two populations. Thus, ComparePop can be used to identify which gene interactions are particular to cancer by identifying the edges in the network that differ significantly between cancer cases and non-cancer cases.

ComparePop compares two populations using a given network. For each edge in the given network, ComparePop computes p-values for the differences of the fit of the edge from the population. The p-values are calculated from the proportions of the two populations which are correctly predicted by rule sets created from the edges. A threshold of the p-values is used to identify the edges which significantly differ with respect to their fit with the given network.

CompareNet and ComparePop are not limited to analyzing networks generated by InduceNet; they may be used to analyze any gene networks. For example, CompareNet could be used to compare an induced network to a manually derived network, such as a known gene network. Thus, CompareNet could be used to compare a network induced from cancer data to a manually generated cancer network. ComparePop could use a manually generated network if the edges in the network were used to create rule sets which could then be given to ComparePop.

### 2.5 Inducing and verifying gene influence networks

We used InduceNet to induce a gene influence network using cases from our combined Beer et al. and Bhattacharjee et al. dataset that were from stage 2, 3, or 4 tumors. We then used ComparePop to identify the edges in this network whose “fit” varied significantly between non-tumor cases from the combined dataset and cases from patients with stage 2, 3, or 4 tumors. This allowed us to identify edges that represented gene interactions that were likely to be more prevalent in stage 2, 3, or 4 tumors, or in non tumors, but not in both.

We verified our findings by comparing the genes in the edges whose fit with the subpopulations varied significantly to a known cancer network [4] and lists of known oncogenes that are available on the web:

1. European Bioinformatics Institute:  
[http://www.ebi.ac.uk/proteome/HUMAN/chromosomes/disease\\_set/](http://www.ebi.ac.uk/proteome/HUMAN/chromosomes/disease_set/)
2. Walter and Eliza Hall Institute:  
<http://gdb.wehi.edu.au/gdbreports/>
3. The Waldman Group:  
<http://cc.ucsf.edu/people/waldman/GENES/completeness.html>
4. CancerIndex:  
<http://www.cancerindex.org/geneweb/clink30.htm>
5. University of Texas Southwestern Medical Center:  
<http://spore.swmed.edu/Luc++Lung+Cancer+genes.xls>
6. University Hospital of Poitiers:  
[http://www.infobiogen.fr/services/chromcancer/Indexbychrom/idxcg\\_X.html](http://www.infobiogen.fr/services/chromcancer/Indexbychrom/idxcg_X.html)
7. Laboratory of Human Genetics, Department of Biology, University of Padova, Italy, has a list of cancer related genes:  
[http://telethon.bio.unipd.it/bioinfo/HGXP\\_151/index.html](http://telethon.bio.unipd.it/bioinfo/HGXP_151/index.html)

### 3. Results

Figure 1 presents the network InduceNet generated from the stage 2, 3, and 4 cases from our combined dataset. Note that many of the edges probably represent gene interactions that might occur in non-tumor cases. With only the information provided by stage 2, 3, and 4 cases, one cannot tell if the interactions would also occur in non-tumor cases. Nor could one distinguish between normal and cancerous interactions by generating a network from all of the cases.

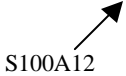
Figure 2 presents the results of using ComparePop to use the network presented in Figure 1 to compare the gene interactions from non-tumor cases in the combined dataset to stage 2, 3, and 4 tumor cases in the combined dataset.

When we examined the genes identified in Figure 2 as possibly being involved cancerous interactions to see if they were already associated with cancer, we found several of the genes in Figure 2 to belong to groups of genes associated with cancer:

- GAS6 and LAMB1 are in the AKT group of genes
- FGF belongs to the WNT group of genes
- PF4 is a member of the TGF-B group of genes
- IL11 is a member of the MEK group of genes

We also found several substructures contained in the networks given in Figures 1 and 2 that are similar (structurally and ontologically) to some of the structures of the “cancer pathway” presented in [4]:

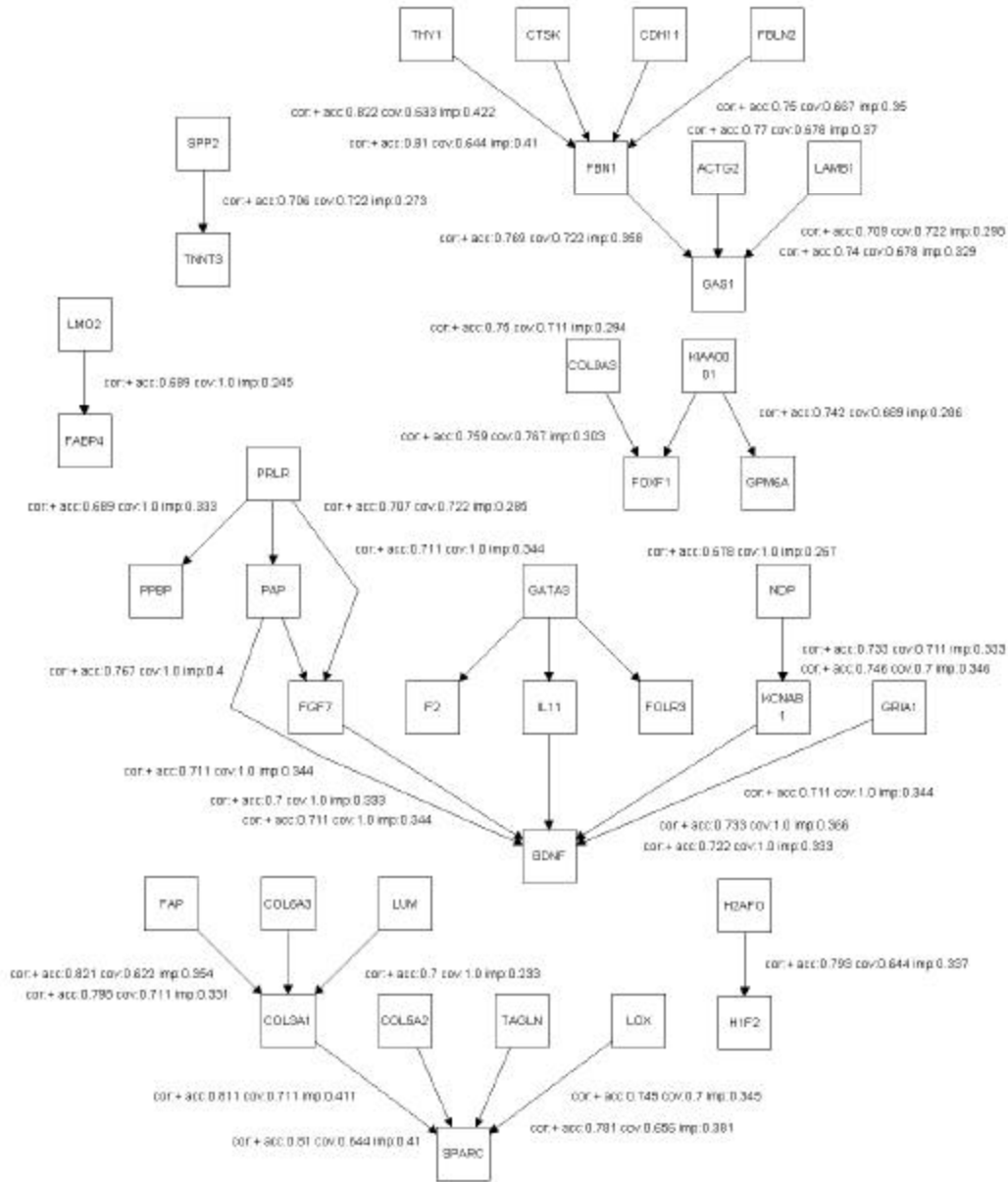
- PF4 ← FGF7 → CD36
- FABP4 ← LMO2 → FOXF1? ← COL9A3
- GAS1 and LAMB1 → SH3GL2 ← ABCC2
- FBLN2 ? → CDH11 ← CTSK? → SPARC ← LUM?  
→ COL9A3 ← GAS6

- GATA → IL11 → SPP2 → TNNT  


We also found LUM, CTSK, FOXF1, and FBLN2 to be known cancer genes. We suspect that a more detailed comparison of our networks to known oncogenes and cancer pathways using ontologies such as the Gene Ontology Consortium’s gene ontology will provide support for the additional genes from Figures 1 and 2.

#### 4. Discussion and Conclusion

We have shown that a microarray database created by merging two existing databases may be used to discover useful patterns. We have also provided methods for using rule induction to induce gene influence networks and provided methods for analyzing gene networks. Our results show that these methods may be used to find differences between the interactions of genes in non-tumor cases and the interactions of genes in stage 2, 3, and 4 cases.



**Figure 1. Gene influence network generated by InduceNet from the combined Beer et al. and Bhattacharjee et al. datasets. The names in the boxes represent genes, and the arrows indicate the direction of influence. The numbers to the side of the arrows indicate, in order, the correlation, positive (+) or minus (-), and the accuracy, coverage, and improvement over the majority class of the rule set which was used as the basis for creating the edge.**

Our current research focus is to extend our program to use ontological data. We are developing methods for using ontological data and models of gene expression and cancer to automatically evaluate and explain the results of our programs and propose hypotheses for surprising findings.

## 5. ACKNOWLEDGMENTS

We give our thanks to Ting Chen, Jianping Zhou, and Jarred McCaffrey, members of our research group that are not listed as authors, but have aided us from time to time. We also thank others

that we have discussed this research with for their helpful suggestions.

## 6. REFERENCES

- [1] Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D., Lizyness, M., Kuick, R., Hayasaka, S., Taylor, J., Iannettoni, M., Orringer, M., and Hanash, S. (2002), Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma, *Nature* 8(8): 816 – 824.

