Learning a Bayesian Network for Melanoma pathways

Radhika Saraf
Department of Electrical and Computer Engineering
Texas A&M University
College Station
saraf.radhika@tamu.edu

Abstract

Melanoma cell signalling pathways have been modelled successfully using Bayesian networks in the past. Cell signalling pathways expose the causal relationships between various genes involved in cell proliferation, death and survival. This work tries to emulate the modelling technique described in *Koller et al.* to learn a directed graph and its parameters from the data. The data set contains expression values of the genes involved in malignant melanoma cancer cells. Once the relationship between the cells has been learned, inference is performed to see the effect of certain genes on others. The goal is to investigate the impact of certain candidate genes on the overall cell proliferation and death and to deliberate if these genes are instrumental in the pathway.

1 Introduction

Melanoma is one of the most prevalent and aggressive forms of skin cancer. In tumor cells, the cell survival mechanism is hijacked by the mutated genes and exploited to counter medical treatment. Metastatic melanoma cells are known to develop resistance to most of the commonly used drugs and therapy.

The human body reacts to threats by relying on its immune system and by appropriate functioning of the cellular signalling pathways. Typically, it will try to battle cancer by inducing programmed cell death or apoptosis in tumor cells, and it will also simultaneously suppress their cell survival mechanisms. Abnormalities in cell cycle control are a characteristic of cancer, and this is accompanied by uncontrolled growth. Typically, most cancer cells deactivate the pathways to apoptosis and simultaneously heighten the effects of the cell proliferation and growth pathways.

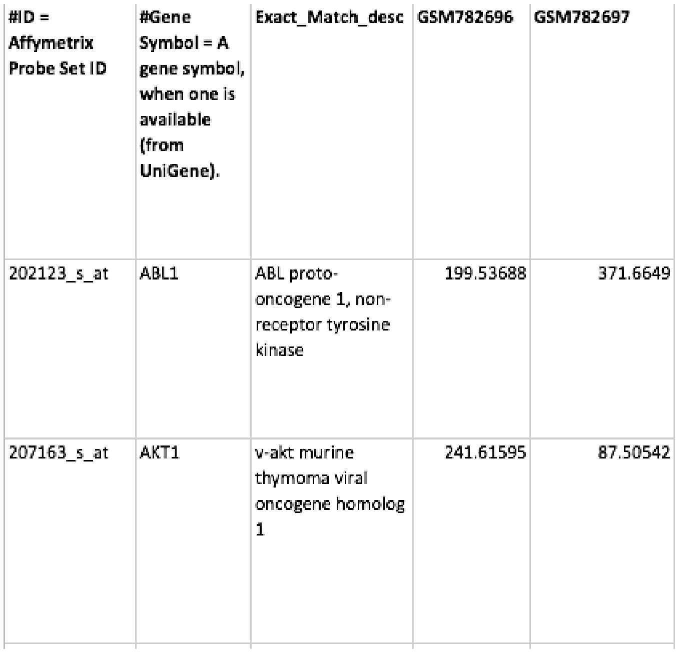
The overall goal of learning the structure is to investigate the genes instrumental in inducing resistance to cell death in melanoma. The various gene interactions in melanoma can be represented by biological pathways such as PI3K/AKT/mTOR and Ras/Raf/MAPK. These cell survival pathways are governed by EGF and IGF, also referred to as growth factors. **The structure learned in this work will be validated by confirming the effect of EGF and IGF on proliferation genes CDK4 and CDK7.** There are certain known mutations in melanoma - NRAS, BRAF, TP53 and PTEN. **The model should quantify their ability to stop cell death.**

After validation, the model shall be used to verify the hypothesis that the gene STAT3 should affect the cell death significantly.

2 Dataset

The National Center for Biotechnology Information provides the values of gene expressions from a series of experiments on melanoma cell lines. GSE31534 is a gene expression profile for A375 melanoma cells after 45 functionally important molecules were knocked down using siRNA (silencing RNA) [3]. **The dataset contains 50,000 genes, of which 44 genes are of interest.**

ECEN 689 Advanced Probabilistic Graphical Models (NIPS 2017), Texas A&M University, College Station.



#ID = Affymetrix Probe Set ID	#Gene Symbol = A gene symbol, when one is available (from UniGene).	Exact_Match_desc	GSM782696	GSM782697
202123_s_at	ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase	199.53688	371.6649
207163_s_at	AKT1	v-akt murine thymoma viral oncogene homolog 1	241.61595	87.50542

Figure 1: Dataset

In [2], the dataset has been used to learn a Bayesian network with around 50,000 nodes. This is clearly a computationally intensive task and to alleviate the cost, the paper describes a method of dividing the dataset into smaller groups and learning subnetworks in parallel. This work looks at one such subnetwork. The final dataset has 77 genes, since the silenced genes are included to reduce the spurious correlations that might occur due to these hidden nodes.

The Figure 1 shows a sample of the dataset. The GSMXXXXX stands for an experiment where one of the genes is silenced using siRNA. There are a total of 51 such samples of each of the 77 genes.

The data preprocessing is described in Figure 2. The value of the gene expression level stands for the relative abundance of that gene. It indicates how much of the gene has been transcribed or activated. A low value means that the gene has been inhibited or inactive.

The project requires only the knowledge of whether a gene is ON or OFF. The method of k-means is used to binarize this data. The two centroids obtained imply that the ON value of a gene fluctuates around one centroid and the OFF value of the gene belongs to the cluster associated with the other cluster. The final output is a 77×51 matrix of 0s and 1s indicating which gene was ON or OFF for which experiment.

3 Model

Gene regulatory networks are usually directed. A Bayesian network should serve as a faithful representation to describe the causal relationship between the genes. The working assumption is that the true cell signalling pathway does not contain any cycles.

Additionally, given the work in [2], Bayesian networks have proven to be good models for Melanoma pathways. The pathways under investigation are PI3K/AKT/mTOR and Ras/Raf/MAPK and there are known not to contain cycles.

The nodes represent the genes (segments of DNA that contain hereditary information) and the edges represent activation or inhibition of children genes based on the status of the parent genes. The probabilistic graphical model tries to model how the transcription (information transfer) of parent genes affects that of the children (downstream effects).

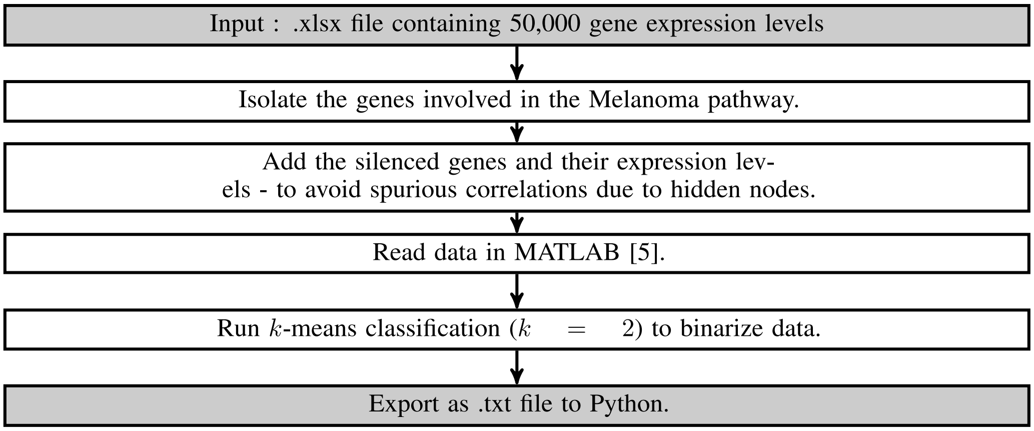

Input : .xlsx file containing 50,000 gene expression levels
Isolate the genes involved in the Melanoma pathway.
Add the silenced genes and their expression levels - to avoid spurious correlations due to hidden nodes.
Read data in MATLAB [5].
Run k-means classification ($k = 2$) to binarize data.
Export as .txt file to Python.

Figure 2: Flowchart for Data Pre-processing

4 Training

The project follows the procedure of finding the class PDAG as described in [1] and in class.

4.1 Finding the P-map

1. Predict skeleton H
   - Check all possible pairs and whether they should be connected.
     $P(X) = X_i \perp X_j | \underline{U}$ then X_i and X_j cannot be directly connected
     Limit largest in-degree
   - $\underline{U}_{X,Z}$ is the witness of independence
   - $X \perp Z | \underline{U}_{X,Z}$ -then H does not contain the direct edge $X - Z$
2. Check potential immoralities
   (a) $Y \in \underline{U}_{X,Z}$ - NOT immorality
   (b) $Y \notin \underline{U}_{X,Z}$ - immorality
   - If the evidence cannot contain Y if $X - Y - Z$ is an immorality.

The witness of independence is limited to 2, in order to try and curb the in-degree of the predicted structure to 2.

The test for conditional independence is done using the χ^2 statistic as the deviance measure $d_{X,Y}(D)$.

$$d_{X,Y}(D) = \sum_{\forall x,y} \frac{(M[x,y] - M\hat{p}(x)\hat{p}(y))^2}{M\hat{p}(x)\hat{p}(y)} \tag{1}$$

This is done using the in-built function in Python package $pgmpy$ [6] in the $ConstrainBasedEstimator$ class using the function $test_conditional_independence$. However, for this model, the run time was 3 hours for each p-value and did not yield satisfactory results. This could be due to the large number of nodes and relatively small number of samples M of each node.

The next test used was conditional mutual information as the measure.

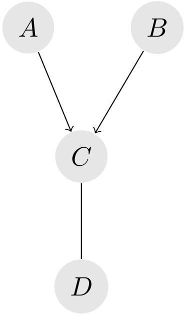
$$I(X;Y|Z) = \sum_{\forall z} p(z) \sum_{\forall x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{2}$$

4.2 Class PDAG

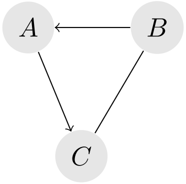
Rule propagation to avoid cycles and assign direction to undirected edges.

1. Direct the edge in a way that no new immoralities are created

2. Direct the edge in a way as to avoid cycles
3. For more than one undirected edge, check how the direction of one affects the others.



The edge must be $C \rightarrow D$, otherwise there will be an extra immorality.



The edge must be $B \rightarrow C$, otherwise there will be a cycle.

4.3 Bayesian Parameter Estimation

Assuming that the prior distribution $P(\theta)$ satisfies global and local parameter independence,

$$P(\theta|D) = \prod_{i} \prod_{Pa(X_i)} P(\theta_{X_i|Pa(X_i)}|D), \tag{3}$$

where D is the dataset given for X_i.

A set of Dirichlet priors is used. The data is binary and the size of each prior will then be 2. If $P(\theta_{X|U})$ is a Dirichlet prior with hyperparameters $[\alpha_{x^1|u}, \alpha_{x^0|u}]$, where $U = Pa(X)$, then the posterior is a Dirichlet distribution with hyper parameters $[\alpha_{x^1|u} + M[x^1, u], \alpha_{x^0|u}] + M[x^0, u]$.

$$P(X_i[M+1] = x_i \big| U[M+1] = u), D = \frac{\alpha_{x_i|u} + M[x_i, u]}{\sum_i \alpha_{x_i|u} + M[x_i, u]} \tag{4}$$

The conditional probabilities are calculated using the baove definition of the posterior.

4.4 Score Optimization

The goal was to maximize the BIC score and present the best model.

$$score_{BIC}(G) = \log L(\hat{\theta_G} : D) - \frac{\log M}{2} Dim[G], \tag{5}$$

where $Dim[G]$ is the number of independent parameters in the model or the degree of freedom of the model. The number of independent parameters depends on the predicted graph. In order to predict the graph, a large computational time was required. This work does not perform the actual optimization of the score, instead a model is chosen as a result of the trial and error method.

5 Evaluation

PI3K/AKT/mTOR and Ras/Raf/MAPK are described in the KEGG database [4] for Melanoma as shown in Figure 4. This true graph is plotted in MATLAB for easier reference and comparison in Figure 7.

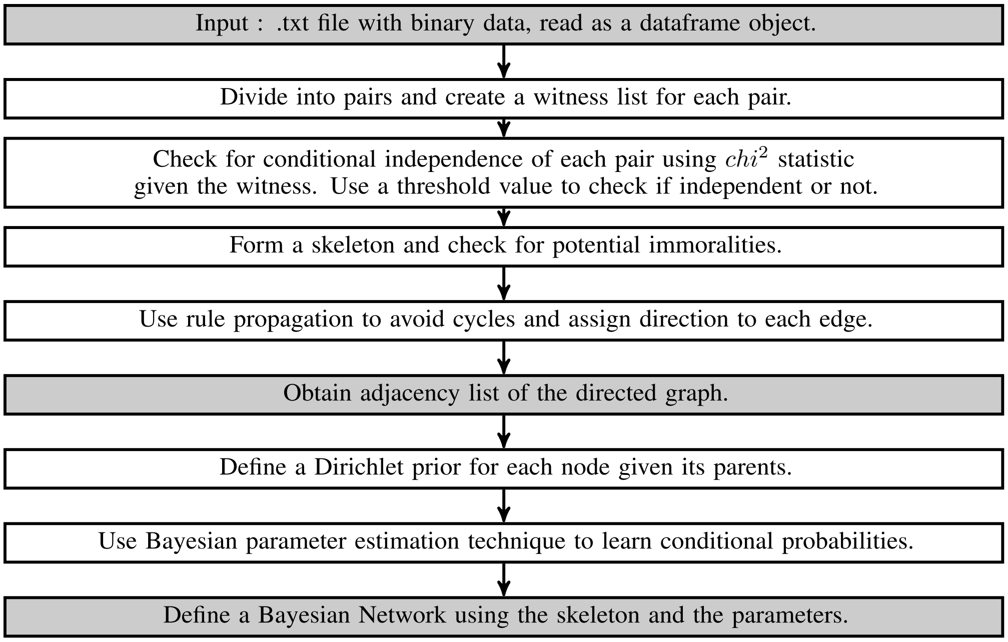

Input : .txt file with binary data, read as a dataframe object.
Divide into pairs and create a witness list for each pair.
Check for conditional independence of each pair using chi^2 statistic given the witness. Use a threshold value to check if independent or not.
Form a skeleton and check for potential immoralities.
Use rule propagation to avoid cycles and assign direction to each edge.
Obtain adjacency list of the directed graph.
Define a Dirichlet prior for each node given its parents.
Use Bayesian parameter estimation technique to learn conditional probabilities.
Define a Bayesian Network using the skeleton and the parameters.

Figure 3: Flowchart for Training

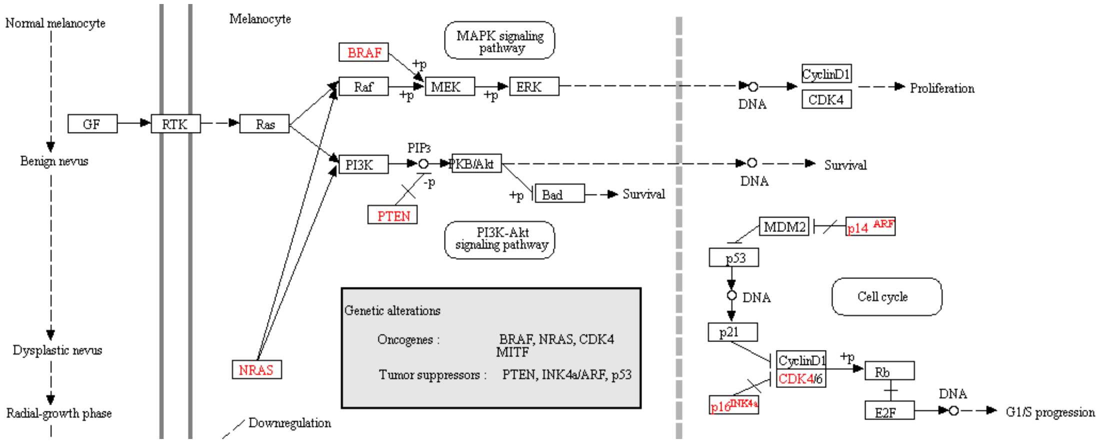


Figure 4: Melanoma Pathways from KEGG database

5.1 Validation

Validation was done using the ROC curve as shown in Figure 6. It is clear that the curve does not show a very good result. The FPR value is 0.0030654, but a closer look shows that

$$('TP', 0, 'FP', 18, 'TN', 5854, 'FN', 57)$$

.

5.2 Inference

Inference was performed using the principals of variable elimination. The 'pgmpy' toolbox was employed to predict the probabilities of certain genes conditioned on others. As shown in Figures 5.2 and 5.2, the probabilities of cell death and proliferation vary with the growth factors and the death ligands. Look at Figure 5.2 in detail, the growth factors are nodes 21, 32 and 48 while the death ligands are 73 and 74. The proliferation factors are 11, 14 and 15. We can see the variation in the probability that cell proliferation is 1 or 0 based on the predicted numbers. Similarly, in Figure 5.2,

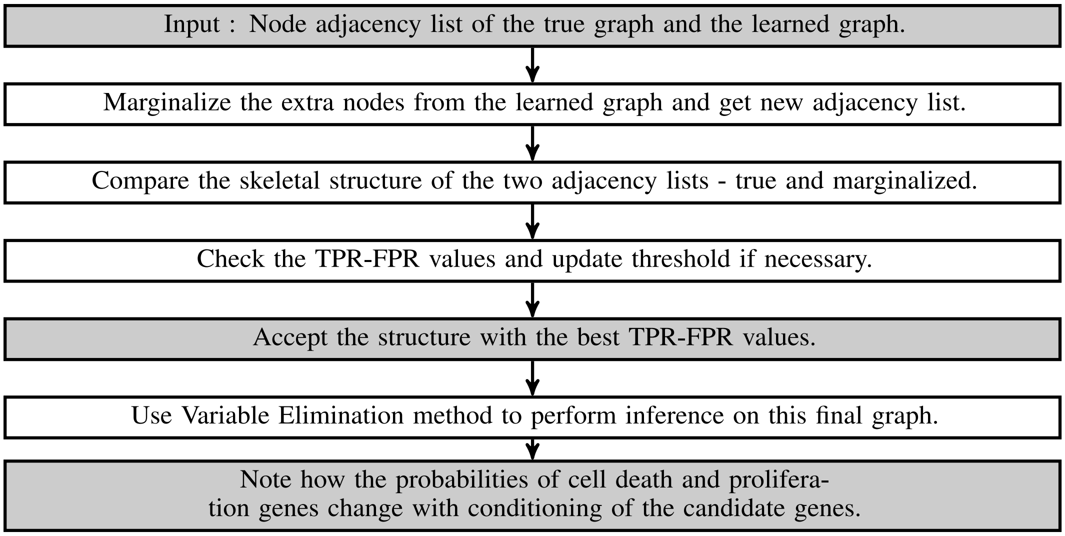

Input : Node adjacency list of the true graph and the learned graph.
Marginalize the extra nodes from the learned graph and get new adjacency list.
Compare the skeletal structure of the two adjacency lists - true and marginalized.
Check the TPR-FPR values and update threshold if necessary.
Accept the structure with the best TPR-FPR values.
Use Variable Elimination method to perform inference on this final graph.
Note how the probabilities of cell death and proliferation genes change with conditioning of the candidate genes.

Figure 5: Flowchart for Validation and Inference

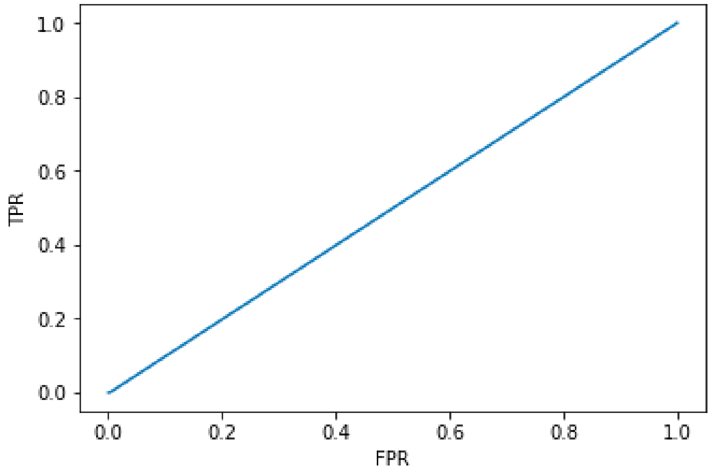
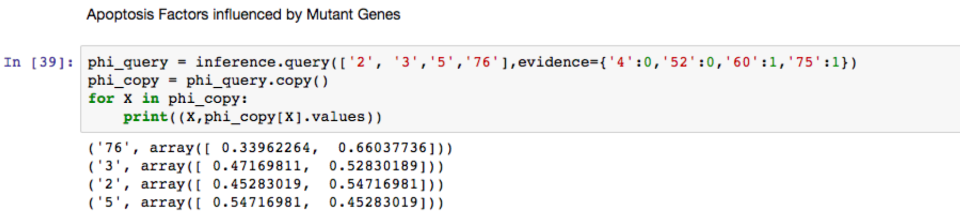


Figure 6: ROC curve

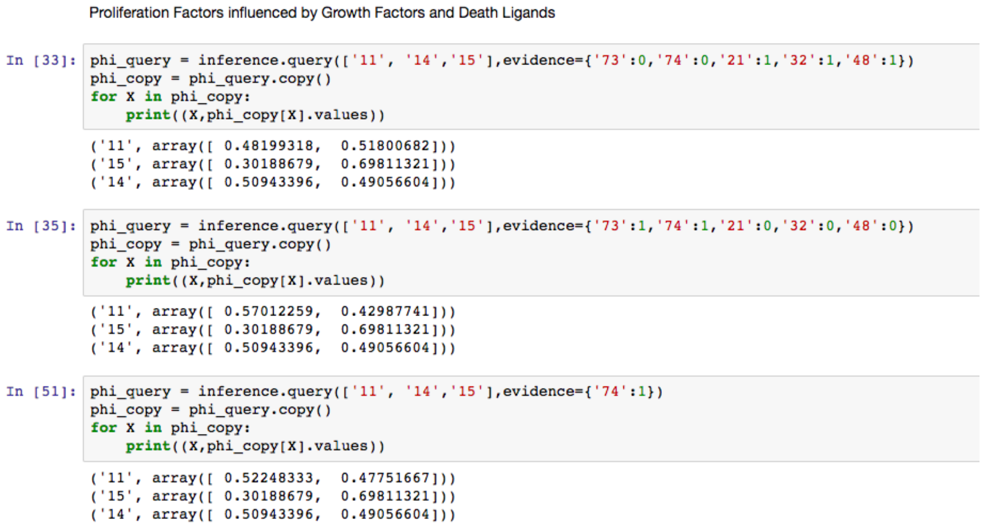
the mutant genes are 4, 52, 60 and 75 and the apoptosis factors are 2, 3, 5 and 76. Here, 2, 3 and 5 cause cell death whereas 76 stops it. The probabilities are accordingly displayed.

To predict whether or not STAT3 is a good candidate, similar inference was performed. As seen in Figure 5.2, the probabilities of apoptosis factors do not change greatly upon conditioning on STAT3, they are similar to the results obtained by the growth factors and death ligands. There is no conclusive proof of STAT3's significance.



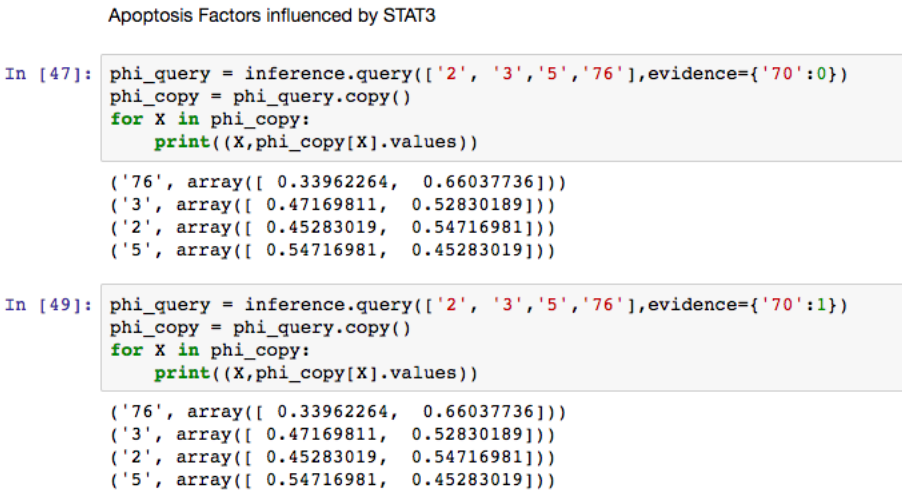
Apoptosis Factors influenced by Mutant Genes

```
In [39]: phi_query = inference.query(['2', '3','5','76'],evidence={'4':0,'52':0,'60':1,'75':1})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('76', array([ 0.33962264,  0.66037736]))
         ('3', array([ 0.47169811,  0.52830189]))
         ('2', array([ 0.45283019,  0.54716981]))
         ('5', array([ 0.54716981,  0.45283019]))
```



Proliferation Factors influenced by Growth Factors and Death Ligands

```
In [33]: phi_query = inference.query(['11', '14','15'],evidence={'73':0,'74':0,'21':1,'32':1,'48':1})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('11', array([ 0.48199318,  0.51800682]))
         ('15', array([ 0.30188679,  0.69811321]))
         ('14', array([ 0.50943396,  0.49056604]))

In [35]: phi_query = inference.query(['11', '14','15'],evidence={'73':1,'74':1,'21':0,'32':0,'48':0})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('11', array([ 0.57012259,  0.42987741]))
         ('15', array([ 0.30188679,  0.69811321]))
         ('14', array([ 0.50943396,  0.49056604]))

In [51]: phi_query = inference.query(['11', '14','15'],evidence={'74':1})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('11', array([ 0.52248333,  0.47751667]))
         ('15', array([ 0.30188679,  0.69811321]))
         ('14', array([ 0.50943396,  0.49056604]))
```



Apoptosis Factors influenced by STAT3

```
In [47]: phi_query = inference.query(['2', '3','5','76'],evidence={'70':0})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('76', array([ 0.33962264,  0.66037736]))
         ('3', array([ 0.47169811,  0.52830189]))
         ('2', array([ 0.45283019,  0.54716981]))
         ('5', array([ 0.54716981,  0.45283019]))

In [49]: phi_query = inference.query(['2', '3','5','76'],evidence={'70':1})
         phi_copy = phi_query.copy()
         for X in phi_copy:
             print((X,phi_copy[X].values))

         ('76', array([ 0.33962264,  0.66037736]))
         ('3', array([ 0.47169811,  0.52830189]))
         ('2', array([ 0.45283019,  0.54716981]))
         ('5', array([ 0.54716981,  0.45283019]))
```

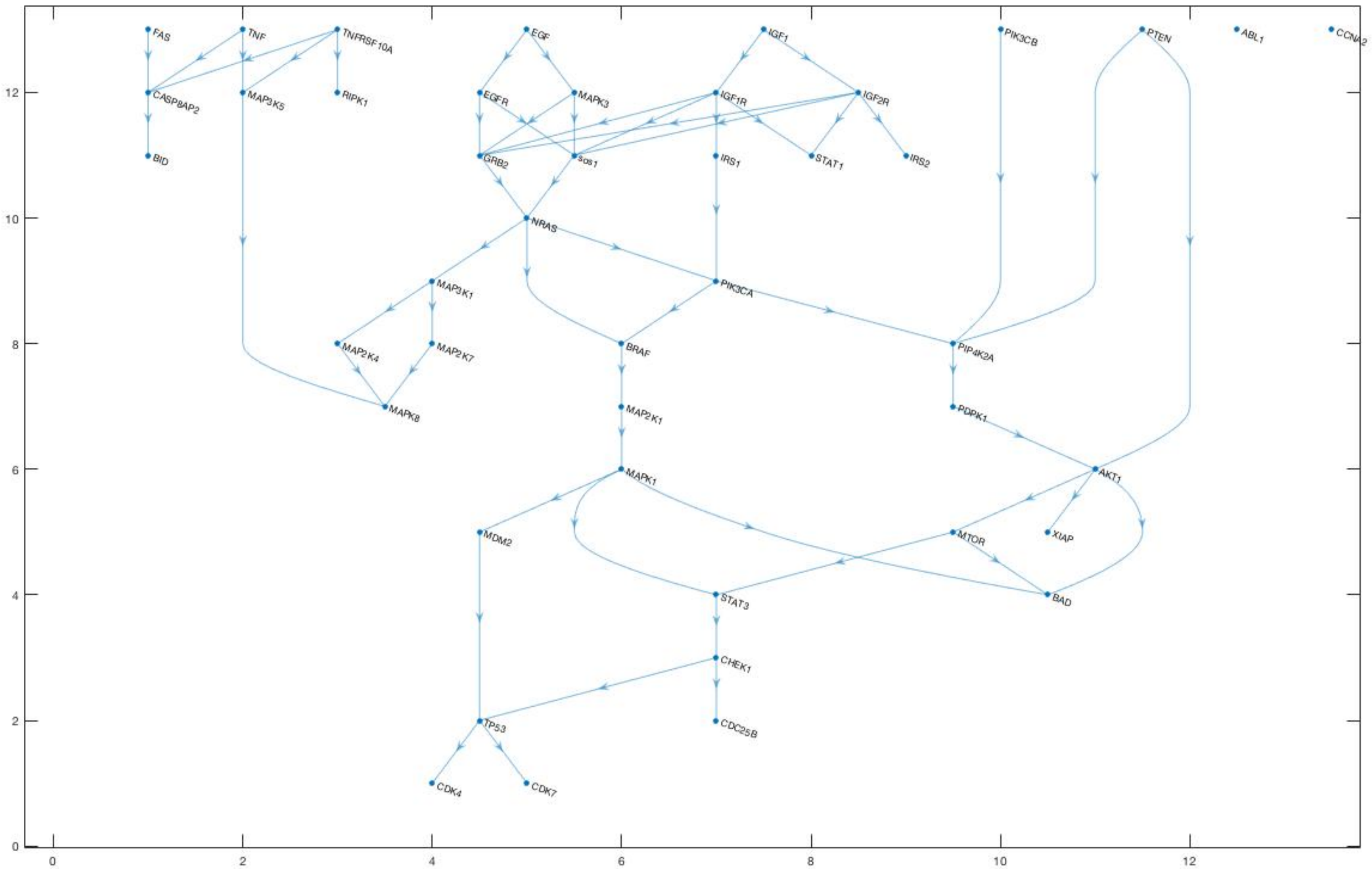


Figure 7: Melanoma Pathways plotted in MATLAB

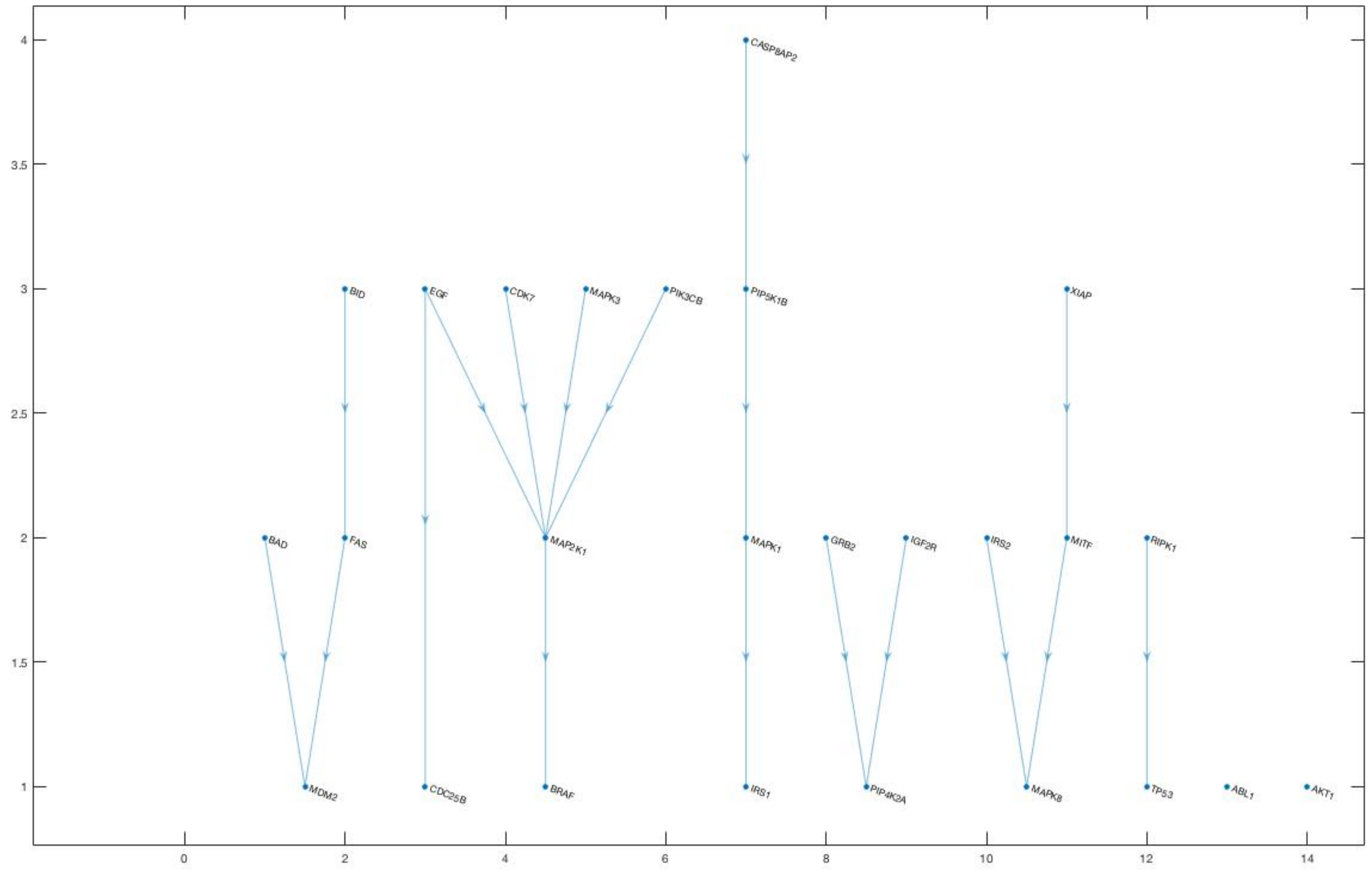


Figure 8: Learned Melanoma Pathways plotted in MATLAB

6 Results and Discussion

The results of the experiment as not very satisfactory. There are no true positives and the inference results are suggestive at best. The validation was not completely satisfactory nor were the inference results.

The possible reasons could be:

- Influence of hidden nodes.
- Lack of data or sufficient samples of each gene.
- Truth was not generated by the same dataset, it was a universally accepted pathway.
- The optimization could not be performed as needed due to computational complexity.
- Data was binarized, instead a continous model might be more apt.

7 Observations, Insights, and Future Directions

Future work would include trying to reduce the number of nodes, and trying for different thresholds and priors. New datasets could possibly be incorporated to make a more detailed dataset.

Acknowledgments

The author would like to thank Prof Byung-Jun Yoon for his continued guidance and support throughout the project. He lent a patient ear to the queries and provided invaluable insights throughout the course duration.

References

[1] Koller. D & Friedman. N (2009) *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning.* MIT Press, Cambridge.

[2] Wang. L, Hurley. DG, Watkins. W, Araki. H, Tamada. Y, Muthukaruppan. A, Ranjard. L, Derkac. E, Imoto. S, Miyano. S, Crampin. EJ, & Print. CG.(2012) Cell cycle gene networks are associated with melanoma prognosis. *PLoS One 7(4)*, e34247.

[3] NCBI GEO database GSE31534 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31534`

[4] Kanehisa M, Sato Y, Kawashima M, Furumichi M, & Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research. 2016;44(Database issue)* pp. :D457-D462.

[5] MATLAB and Statistics Toolbox Release 2012b, *The MathWorks, Inc.*, Natick, Massachusetts, United States

[6] Python toolboxes `https://github.com/pgmpy`, `http://scikit-learn.org/stable/`