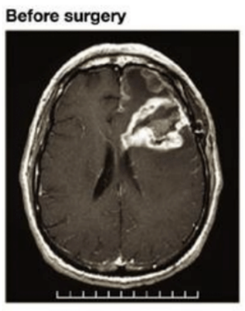
A BAYESIAN FRAMEWORK THAT INTEGRATES TRANSCRIPTOMICS AND VARIABLE SELECTION TO PREDICT RISK GENES FOR GLIOBLASTOMA

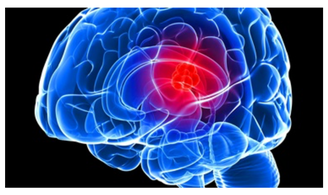
Rohan Ahluwalia. Sergio Branciamore Ph.D, Andrei Rodin Ph.D
Department of Diabetes Complications and Metabolism, City of Hope, Duarte CA

INTRODUCTION & OBJECTIVE

Glioblastoma Multiforme (GBM) is the most common type of Central Nervous System Tumor.

- The prognosis for patients who develop GBM is bleak, with average survival after diagnosis ranging from 12-16 months
- It is a very fast growing tumor that spreads creating pressure in the brain and various painful symptoms
- Standard treatment includes surgical debulking followed by TMZ (Temozolomide) and IR
- Major Challenge in patient treatment is the recurrence of these tumors after the initial therapeutics have been deemed successful





There are three **objectives** of this project:

1.Identify risk genes for glioblastoma with high confidence

2.Develop an effective method for Bayesian network analysis of temporal data

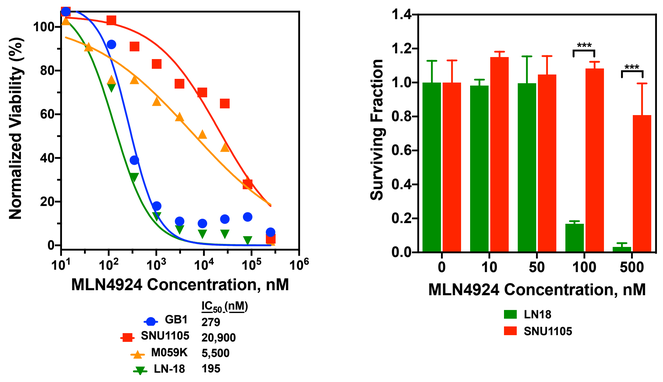
3.Determine the biological significance of risk genes predicted

PREVIOUS EXPERIMENT & TEMPORAL DATA

There was a transcriptome dataset provided by Translational Genomics Research Institute (TGen) which had a genomic database with multiple features.

Previous Experiment

A previous experiment was conducted to identify is Pevonedistat could be a targeted therapy in GBM. By conducting this experiment there was four long-term glioma cell lines with a differential response to Pevonedistat – LN18, GB1, M059K, and SNU1105.



This figure shows that the cell lines had a differential response to MLN4924, or Pevonedistat

Based upon these results, a dataset was provided for better understanding of the differential response

Dataset

Target Variable 1 *Cell Line*	Target Variable 2 *Drug Concentration*	Target Variable 3 *Time*	Gene Data *~36000 genes*
Discrete Data (4 Cell Lines)	Discrete Data (4 concentrations)	Discrete Data & Longitudinal Aspect	Continuous Data

In addition to the mixture of discrete and continuous data, there was a longitudinal aspect to the data.

- Longitudinal data consist of repeated measurements of some variables which describe a process over time
- A temporal dataset is a dataset with built-in time aspects which changes the data from 2D to 3D, with one of the axis being time
- Currently, there are **no well-established methods** for analyzing **temporal** data, especially in the Bayesian network environment

BAYESIAN NETWORK & TEMPORAL DISCRETIZATION

A Bayesian Network (BN) is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- The reconstruction of the network is a 2-stage process
  - Model Selection (Search of a network structure that best fits the data)
  - Probabilistic Inference Propagation given the network structure

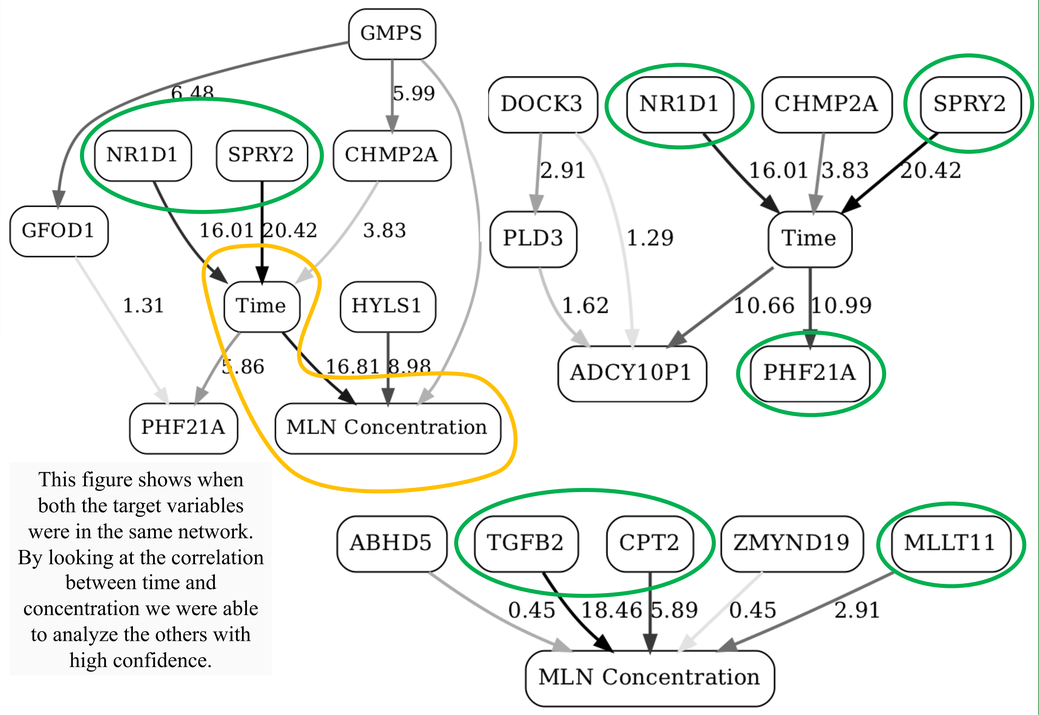
The process has three main stages:

- Starting with an empty network and force feed certain connections based on conditional entropy
- Now, these connections are then scored and analyzed based upon entropy
- Based upon these scores, further connections are made and the nodes are either added or dropped

To analyze the temporal data, we were able to create a method to discretize the data that made biological sense to determine the relationship between time and the other target variables.

RESULTS

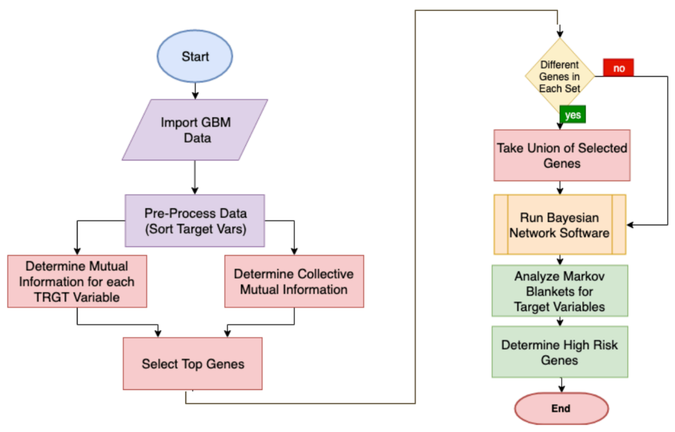
The results from the Bayesian network analysis method provided multiple high risk genes and interesting correlations between target variables



This figure shows when both the target variables were in the same network. By looking at the correlation between time and concentration we were able to analyze the others with high confidence.

DEVELOPMENT OF ALGORITHM

The proposed algorithm consisted of five distinct steps and is highlighted in the following flowchart.



1. Preprocess Dataset – Preprocess dataset and set the target variables for analysis
2. Determine Mutual Information – Reduce the size of the dataset to 1% of original
3. Discretization Method – Discretize the time variable to analyze properly
4. Run Bayesian Network – reconstruct a network based on conditional entropy
5. Analyze Markov Blankets & Determine High Risk Genes

MUTUAL INFORMATION

Mutual information (MI) is a measure of the mutual dependence between two variables. This quantifies the amount of information shared between two variables.

$$\hat{I}(X;Y) = \sum_x \sum_y p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

The problem with the dataset is that there is a mixture of continuous and discrete data, which makes estimating the mutual information with the equation above extremely difficult.

Therefore, a new method was implemented using k-nearest neighbors to estimate the mutual information.

MI is the average of the logarithm of Radon-Nikodym derivative, so compute this for each value *i* and take the empirical average

- When the point is discrete (distance of k-nearest neighbor = 0), use the plug-in estimator for Radon-Nikodym derivative

$$\hat{I}(X;Y) = \sum_{i=1}^{n} \log\left(\frac{dP(x,y)}{dP(x)P(y)}\right)_{(x_i,y_i)}$$

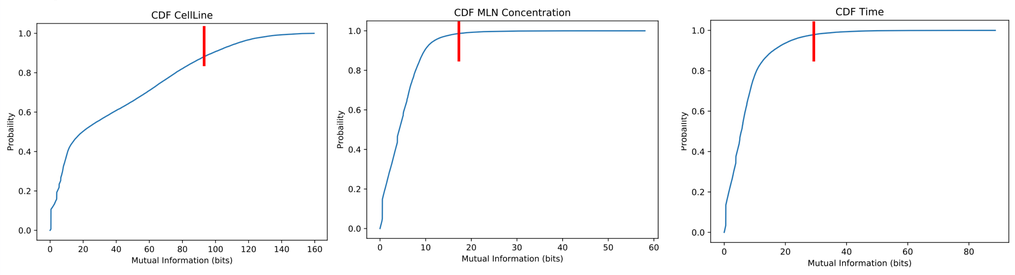
- If the distance of the k-nearest neighbor is not 0, then it must either be mixed or continuous, then use the estimator below

$$\hat{I} = \psi(k) + \log(N) - \log(n_x, i+1) - \log(n_y, i+1)$$

Where ψ is the digamma function.

VARIABLE SELECTION

After the mutual information was estimated, the variables were selected so that the most impactful genes would be analyzed.



RISK GENES

Based upon the proposed method, there were numerous risk genes that were predicted with high confidence.

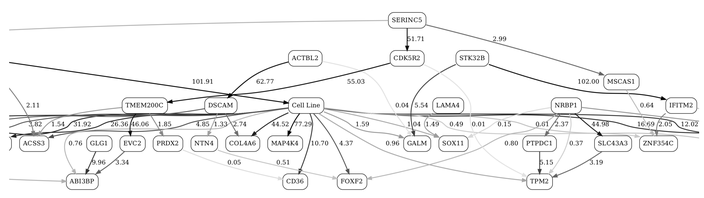
Gene	Description	In Literature
NR1D1	Transcriptional repressor which coordinates circadian rhythm and metabolic pathways in a heme-dependent manner.	Yes
SPRY2	Functions as an antagonist of fibroblast growth factor (FGF) pathways and may negatively modulate respiratory organogenesis	Yes
TGBF2	A secreted protein known as a cytokine that performs many cellular functions and has a vital role during embryonic development	No
CPT2	This protein is in the mitochondrial inner membrane. With other proteins, it oxidizes long-chain fatty acids.	No
HYSL1	Mutations in this gene are associated with hydrolethalus syndrome	No
CLND1	Claudins function as major constituents of the tight junction complexes that regulate the permeability of epithelia	No
MLLT11	Has been shown to be fused with a number of translocation partners in cases of leukemia	No
PPP1R18	Interacts with regulatory subunits that target the enzyme to different cellular locations and change its activity toward specific substrates.	No

By identifying both genes that have shown previous connections to glioblastoma and some novel genes, it provides a good basis to predict with high confidence that these genes impact glioblastoma in various ways.

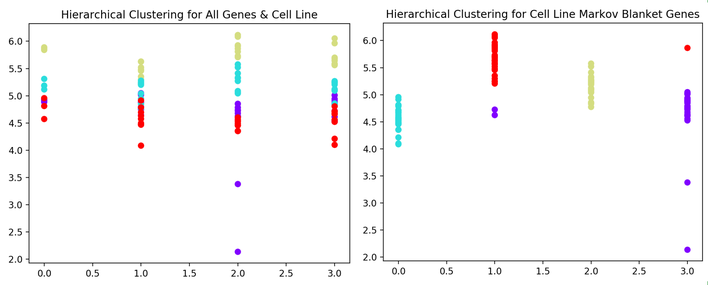
VALIDITY OF RISK GENES

CLUSTERING ALGORITHM

To validate the genes that were selected for the risk genes, a cluster algorithm was used to verify that the important genes were responsible for the changes determined. This cluster algorithm was done only on the cell line feature because it was the most difficult to understand.



Only a part of the blanket is shown for an idea of why the clustering strategy was needed.



These figures show that the genes determined by the proposed algorithm were the impactful genes because they were responsible for the clustering.

Biological Conclusions

Based upon the proposed algorithm and the various method to determine the accuracy of the risk genes predicted, we are confident that the set of risk genes predicted have large impact on glioblastoma.

- The clustering approach supported the claim about the genes that were predicted with respect to the cell line
- The literature research proved some of the relationships of genes which provides a good basis to make confident predictions
- By determining a relationship of time and drug concentration we determined a baseline to make confident predictions for risk genes

Future Work

- Develop a more robust strategy for analyzing the temporal dataset
  - Possibly using time as a feature in the conditional entropy calculation
- Analyze the risk genes selected in a biological study to show that these genes are significant with relation to glioblastoma
- Analyze the dataset using different Bayesian network formats to see the variance in risk genes predicted
- Develop potential therapeutics for glioblastoma based upon the risk genes determined by the algorithm

ACKNOWLEDGMENTS & REFRENCES

Special thanks to Dr. Sergio Branciamore, Dr. Andrei Rodin for guiding me during this project & for City of Hope Diabetes Summer Research Institute for having me this summer. In addition, thanks to TGen for providing the dataset.

References:
[1] Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., . . . Li, B. (2019, April 15). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data.
[2] Branciamore, S., Gogoshin, G., Giulio, M. D., & Rodin, A. (2018). Intrinsic Properties of tRNA Molecules as Deciphered via Bayesian Network and Distribution Divergence Analysis. Life, 8(1), 5. doi:10.3390/life8010005