CS410 Final Project Report

Zhenglun Chen

Introduction

Cancer and its relation to gene mutation have always been a fascinating topic in research these years. In the beginning of this semester, by chance I have acquired a comprehensive and authentic dataset of cancer related gene mutations and its patients. The data covers exact gene that is mutated for each patient and the length of treatment of each patient. In my research, I examined the the correlation between the gene pairs to studied if the mutation of any particular gene pair has significant effect on the survival of a patient.

Related Work

In my research, I incorporated a newly proposed method called hypercontractivity measurement of potential correlation which is proposed by Gao (Link to his paper). Also, I will also examine the well-known measure of pearson's correlation coefficient. The traditional pearson's correlation coefficient tends to perform well on linear dataset, but its performance on nonlinear data is yet to be discussed. In Gao's paper, he has proved that the correctness and accuracy of hypercontractivity measurement in discovering the potential correlations. Therefore, in the process I will compare and contrast these two different work and how they perform on my data set. Other than these two methods, I will also involve fisher's exact test and survival log-rank test to complete my study.

Methods

Before I talks about the details, I'd like to present the overview of the format my dataset to provide a better preview of my project. There are two types of data: 1) Mutation data 2) Survival data. Their format are shown below.

Index	Patient	Gene
0	TCGA-WD-A7RX	79632
1	TCGA-WD-A7RX	761
2	TCGA-WD-A7RX	23043
3	TCGA-WD-A7RX	23040
4	TCGA-WD-A7RX	9675

Index	ID	OS_MONTHS	OS_STATUS
0	TCGA-WD-A7RX	0.69	1
1	TCGA-W5-AA30	35.5	0
2	TCGA-ZD-A8I3	1	1
3	TCGA-W5-AA2R	48.8	0
4	TCGA-4G-AAZO	20	0
5	TCGA-W6-AA0S	11.6	0
6	TCGA-3X-AAV9	0	0
7	TCGA-4G-AAZT	6.77	0

On the left side, it is the patient and gene mutation data. With the first column being patient ID and the second indicates the gene code that got mutated. On the right side, its is the patient and survival data, which "OS_MONTHS" being the time of treatment and "OS_STATUS" being an indicator where 0 means the patient is survives and 1 means the patient failed to survive the cancer.

As the data are organized in the format above, the main steps of my project are the following:

1) Convert data into vector
2) Construct contingency table and perform Fisher Exact Test to filter data
3) Compute pearson's correlation coefficient and hypercontractivity
4) Filter the 50 most uncorrelated pair
5) For each pair filtered construct the table for log-rank test and conclude the result

Now I will go over these steps one by one in detail, step 1-4 only involves patient-gene data.

1) For each cancer gene, I construct an inverted vector. The length of the vector is the number of patient in dataset. For a gene G, the vector of G is denoted as $V_G(a)$ and for every patient a, $V_G(a) = 1$ if the gene G of patient a is mutated (0 otherwise). Most of the genes are only mutated in few patients, I choose to opted out the gene which mutated in less than 10 patients to avoid unnecessary computation.

   Here is a sample look of gene vector. As shown, this is the vector with size 194 because there are 194 patients. For each gene, for example, gene 113146 is mutated in patient 3 and patient 4 (more are not shown in the figure).

Key	Type	Size	Value
10075	list	194	[0, 0, 1, 0, 0, 0, 0, 0, 0, 1, ...]
10178	list	194	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...]
10721	list	194	[0, 0, 0, 0, 0, 0, 1, 0, 0, 1, ...]
10847	list	194	[0, 0, 0, 0, 1, 0, 1, 0, 0, 1, ...]
1108	list	194	[0, 0, 0, 0, 0, 0, 1, 1, 0, 0, ...]
113146	list	194	[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, ...]
114788	list	194	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...]

2) After I yielded the gene vector for each cancer gene , I investigated the correlation between any pair of genes. To do so, I first constructed a contingency table in the format below. Then I used fisher's exact test with left-side p value to test on whether the number of mutation is independent of type of gene in the pair. The left-side p value indicates an alternative hypothesis of Gene 1 is less likely to mutate.

	Gene1	Gene2
Mutated		
Not Mutated		

At this step, I enumerated all the possible pair of combinations in for two genes. After calculating the p value, I decide to omit the pairs with p values greater than 0.3. Because with such large p value, it could easily accept the null hypothesis.

3) From step 2, I have yielded the gene pairs with p value of fisher's exact test less than 0.3. To continue investigating the correlation behind each pair, I decide to compute pearson's correlation coefficient meanwhile also hypercontractivity for each pair. To my surprise, both coefficient performs well on the binary data and they yield same result: both of them tends to approach to 0 when two genes in the pair are statistically independent.
4) After finding the correlation between any pair of genes, I choose 50 most uncorrelated gene pairs to investigate on, that is, the pair with lowest hypercontractivity and lowest absolute pearson's correlation coefficient. The reason behind this is that I want to find out if two genes are really uncorrelated, what would happen if both of them are mutated in 1 patient (i.e. how will the mutation of both genes affect the survival of a patient).
5) From the 50 pairs of genes I got from last step, I construct a survival dataset to perform survival log-rank test for each pair as explained below.
   For each pair of gene denoted as P, let P = (G1,G2) where G1, G2, I labeled each patient in patient survival data as following: if both G1 and G2 are mutated in the patient, the patient is labeled as 2, otherwise labeled as 1. With this label, combining the OS_MONTH and OS_STATUS data, I could perform survival log-rank test. Assume the significance level is 0.05, if the p value of the test goes below 0.05, it means the gene pair that got tested on significantly influences patient's survival if both genes are mutated. Here is a sample file for gene pair (472,55294) for CESC mutation. As illustrated, the patient TCGA-EA-A43B got treated for 25.99 month with both gene 472 and gene 55294 mutated.

Result:

The number of mutations I researched on is 15. (All the files are uploaded into github). For each mutation, I have discovered the gene pairs that has significant effect on patient's survival. The table below is my discovery.

CESC	(120114,58508)	LGG	(23742,1956),(23742,3417),(338,1956),(338,4763),(442444,3417),(442444, 6597),(4626,1956),(5728,4851),(82872,6597)
PAAD	NA	LIHC	(196385, 8289),(2200, 8289),(26960, 53353),(54768, 8289),(54798, 8289),(56171, 8289),(57578, 8289),(65217, 8289),(80144, 8289).
OV	(1293,82872) (2312,82872)	KIRC	(131873, 8242),(23345, 8242),(667, 29072).
SARC	(5925, 7273)	GBM	(10735, 140453),(1278, 7157),/(1756, 7157),(2312, 94025),(388698, 7157),(58508, 7157)
KIRP	(2195, 4583), (29072, 4771), (84033, 4583), (90850, 4583)	PRAD	(113146, 114788),(113146, 140453),(113146, 53353),(113146, 58508),(113146, 7157),(53353, 140453),(79026, 113146),(79026, 114788),(79026, 140453),(79026, 472),(79026, 53353),(79026, 58508),(83872, 140453),(83872, 53353)
TGCT	NA	THCA	NA
KICH	(4588, 7157), (94025, 4588). (94025, 7157)	ACC	NA
CHOL	NA		

Moreover, other than the gene pairs yielded for each location, I also found that pearson's correlation coefficient performs well on binary data as well as hypercontractivity.

Discussion:

The result gene pairs of this research project is yet to be discussed. However, to interpret such result, relevant knowledge in genetic science or related field is needed. Moreover, after yielding the gene pairs, it is possible to go further in this project by proposing a prediction model of patient's survival for different mutations based on the result we yield in the previous section. This project can also be improved by tuning parameters for the methods involved in the process.