DATA201 Project report

Team member: Zhuoheng Chen

Data source:

Gene Database
http://xena.ucsc.edu
TCGA-GBM.htseq_fpkm-uq.tsv:
https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-GBM.htseq_fpkm-uq.tsv.gz
TCGA-GBM Survival
https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-GBM.survival.tsv
Genecode.v22
https://gdc-hub.s3.us-east-1.amazonaws.com/download/gencode.v22.annotation.gene.probeMap
THE GENE ONTOLOGY RESOURCE(GOEnrichment Analysis)
https://www.geneontology.org
KEGG Database
https://www.genome.jp/kegg/kegg2.html

Goal:

Reduce the size of the database and identify which of these genes have the most significant influence on the formation of brain cancer.

Why choose these data

After reading some URL links in the document published in the assignment, I found that I did not find some database that I was very interested in. Then, when I accidentally talked about this topic with my friend, who is a biology student, he provided me with such an inspiration. However, I did not know biology very well, so I hesitated at the beginning. But I was finally convinced by him, so I chose to do such a project. Of course, there are many kinds of cancer, and I chose to do brain cancer and this individual case under brain cancer was just randomly selected. At that time, I read some lists of cancer categories and found that there were some female diseases such as uterine cancer and breast cancer that I, as a man, did not know about. So, I still want to do some cases related to men, so I chose brain cancer, and the database is relatively easy to find, and the data contained is relatively comprehensive, so I decided to start doing this database.

What target I choose

We started out with more than 60,000 data in the database and the initial goal was to reduce the number of data to single or double digits, also identify some of the most significant genes to figure out which genes are responsible for the disease.

At the same time, since the task is written about how your final database and conclusions can help other scientists, it is hoped that finding out the cell sense that has significant factors in the composition of cancer can provide some help to bioscientists, which is also in line with the assignment needs.

Difficulties and techniques

There were a lot of difficulties in this process and the first one was that when I was looking at over 60,000 subsets of data it wasn't clear where I was going to start until I realized that because not every gene is going to be a cancer or in other words the gene expression of a normal gene is going to be different from the gene expression of a cancer cell, The first task was to distinguish cancer cells from normal cells. When I finished the task of distinguishing cell tissues, I found that I could not find any other way to further simplify the data, so I changed another database and decided to complete the analysis of cancer cells from the survival rate of patients.

First, I used two enrichment analysis methods commonly used in biology to screen and identify the data, but these two analysis methods could not help me simplify the amount of data, only tell me the role of these genes and their pathways. Then, after a new round of learning, I chose single-factor Cox regression analysis to screen the variables. Such a method can indeed help me eliminate the data that are not helpful to my final data model, but the amount of data remaining is still very large. Then I chose lasso regression and multi-factor regression model to further screen the data. Finally, I got a satisfactory answer.

Manage to achieve and failed to do

In this task we achieve our initial goal of finding some significant genes that contribute to brain cancer from a huge database. Of course, there are many things that can be done better. For example, in the first database, I believe there must be a better way to filter and reduce the data, but maybe I can't get more information from the first database because my ability and ideas have not reached the corresponding level. Also, the resulting cells can be studied in more depth, such as why does it have a significant factor, and what diseases might it cause.