Survival Analysis of TCGA dataset

Sailendra Pradhananga (sailendra.pradhanaga@scilifelab.se)

Project Report:Algorithms in Bioinformatics

February 14, 2018

1 Introduction

Survival analysis are statistical toolkits that are used to analyzing the outcome of variables based on time period of event occurrence. The event could be either be death, outbreak of disease or failure of automobile parts which are then followed specified time period in days, weeks or years. The major part of the these type of data involves two kinds of variables: first is the time to event and second whether the event have occurred or not. Based on different methods such as non parametric model such Kaplan Meier method or regression model such as cox proportional hazards two function, survival function and hazard function. The survival function gives the probability for each time-points of surviving up to that events while hazard function whether the event has occurred or not given the individual has survived or not. The main aim of these type of analysis is find the relationship between the survival time and variables under study such as drugs, gene expression, treatment either with different covariates such as age, gender, race.

In the current project, we are provided a liver cancer patients of different races dataset survival time and gene expression dataset of various grades. The main aim was to model the data with survival model in order find the potential prognostic marker for survival of patients using Cox Proportional Hazard models. Additionally, using the multiple hypothesis correction we predicted 4698 gene transcripts at 0.05 FDR.

2 Method

2.1 TCGA Dataset

The TCGA dataset for the analysis was download from the (http://kaell.org/files/survivalLIHC.txt) site. As already stated in the project guidelines, these are the cancer patient expression data from the sequenced liver tumor. Additionally there are few

other features implication each cancer patient ethnicity, sex and cancer types. These data were first read into python notebook as pandas dataframe as done in jupyter notebook. Additional exploratory analysis was done in the notebook itself. The dataset consists of both categorical features and numerical data of gene expression dataset.

2.2 Preprocessing of Dataset

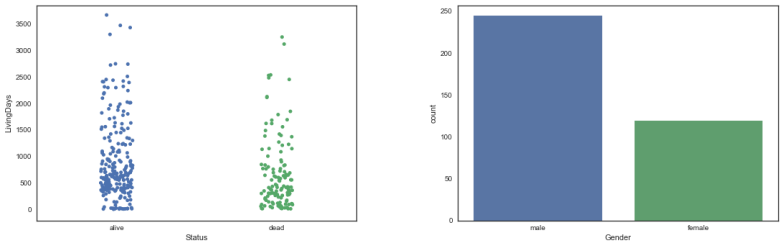
In the survival analysis, the different categorical features were provided of liver cancer patients. These categorical features includes Gender, Race, Stage , Status, Age and LivingDays of different cancer patients. These categorical features were then coded into different numerical codes for Originally, the dataset consists of gene expression profile os $19,571$. Furthermore, only those gene expression data were taken which had at least an expression in one of the patients.

2.3 Cox proportional Hazard model

3 Result

3.1 TCGA Liver cancer charcterstics

The data consists of 6 catagorical features as



(a) A gull (b) A tiger

Figure 1: Pictures of animals

3.2

3.3