

Survival Analysis of TCGA dataset

Sailendra Pradhananga (sailendra.pradhananga@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

2.2 Preprocessing of Dataset

2.3 Cox proportional Hazard model

3 Result