

TCGA Liver cancer survival analysis of expressed gene transcript

Sailendra Pradhananga (sailendra.pradhananga@scilifelab.se)

Project Report: Algorithms in Bioinformatics

KTH Royal Institute Of Technology

February 16, 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 survival analysis dataset consists of two kinds of response variables: first, the time to event and second the information on event occurrences. Various parametric and semi/non - parametric methods are developed to evaluate time based response of event. Kaplan Meier method or regression model such as cox proportional hazards evaluates two functions that is survival function and hazard function based on these response variables with other categorical and numerical dependent variables. 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 survival. 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 or independent on different covariates such as age, gender, race.

In the current project, using publicly available high-throughput RNA expression of liver cancer tumor sample, subsequent phenotype characteristic and time of survival information, we aimed at evaluating survival analysis of each gene transcript. The raw data were subsequently preprocessed, fitted into cox proportional regression hazard model and each gene transcript was evaluated for potential prognos-

tic marker for survival of patients at 5% and 1% false discovery rate.

2 Material and Method

2.1 Acquisition of raw TCGA Dataset

The TCGA dataset for the analysis was download from <http://kaell.org/files/survivalLIHC.txt> website. As already stated in the project guidelines, these are liver cancer patient expression data from the sequenced liver tumor. The dataset consists of both categorical features and numerical data of gene expression dataset. Therefore, there are few additional features of each cancer patient including ethnicity, sex, age, liver cancer sub-types and survival time. This raw dataset were first read into python notebook as pandas data frame as done in jupyter notebook, https://github.com/sailepradh/prj_algo_bioinfo/blob/master/notebook/TCGA_project.ipynb. Additional exploratory analysis was done in the notebook itself.

2.2 Preprocessing of Dataset

The provided raw dataset consists of detailed information of patient phenotypes, however, survival analysis model are unable to process these information. Hence these information needed to be categorically coded before proceeding into actual analysis. Phenotypic features of cancer patient sample such as gender, race, status, liver cancer subtypes were coded into different codes. Furthermore, gene transcripts

which lacked expression data in atleast one of patients were removed from further analysis as these feature will unable to provide prognostic marker information.

2.3 Cox proportional Hazard model

Cox proportional hazards model are regression model used to relate the survival time with one or more predictor catagorical and quantitative variablees. As our goal in current project is test expression of each transcript as prognostic marker of survival, Cox proportional hazard model aptly fits the requirement. Each gene transcript expression values are modeled against the survival time and survival status for patient sample. To this end, we specifically used the Cox regression hazard model implementation in python from lifelines [1] module. Based on proportional hazard model we get a hazard function, survival probability and p-value estimates of gene transcript as prognostic marker. However we found that not all the gene transcripts were fitted into the Cox survival model ,thus we excluded to get final list of pvalues for each transcripts.

2.4 Multiple testing correction

In order to control the false postive rate due to multiple feature testing, we implemented Storey and Tibshirani [2] q value estimates based on cox modeled p-value of gene transcript. As discussed in earlier exercises we have implemented multiple testing correction Furthermore, we estimated $\hat{\pi}_0$ from differnt λ values in order to make q value estimates of i th significant feature. These implementation were made on python and can be observed in associated notebook.

3 Result

3.1 TCGA Liver Cancer patient phenotype

The cancer cohort consists of different phenotypic features such liver cancer types, age, gender, which are further summarized in table 1 . The major of

samples were white male with the stage i liver cancer subtype. Importantly, as we are interested in survival analysis, the important dependent variables are the status of each patient and the living days which is then fed into the survival analysis. To end this , we observed that average living days for alive patient cohort is significantly different: patients that were dead lived 655.21 while living had living days of 896.80 at $p < 0.01$ as figure 1

Summary statistics of TCGA liver cancer cohorts	
Gender	
male	245
female	119
Race	
white	182
asian	154
black or african american	17
not reported	10
american indian or alaska native	1
Liver cancer stage	
stage i	170
stage ii	83
stage iiia	63
not reported	24
stage iiic	9
stage iiib	8
stage iii	3
stage ivb	2
stage iv	1
stage iva	1
Status	
alive	234
dead	130

Table 1: Summary Statistics of TCGA liver cancer cohort

3.2 Gene expression survival analysis with Cox regression model

The survival analysis requires the time period for the event to occur (in our case death). Thus , in liver cancer cohort variable living days was taken as time variable with status as hazard event. However,

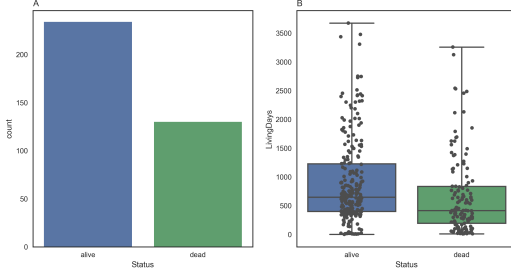


Figure 1: A) Number of alive and dead patients in the liver cancer cohort, the number of alive patients are nearly twice as many as dead patients. B) Distribution of living days in two status group , alive and dead.

other catagorical variables had *NANs* data so these were not taken into consideration while building Cox regression hazard model implementation in lifeline module in python. Each transcript was taken as dependent variables in the model which gave us survival probabilities and cumulative hazard estimates for each transcript as shown in the figure 2. However not all the gene transcript ran into the model as these gene expression were spares and had missing expression data. These 130 features were excluded from the current analysis. The cox regression model was run in 19219 gene transcripts completely which gave us hazard coefficient and survival probability based on expression profile of each gene transcript. These estimates provided the p-values as the measure of the significance of gene expression as prognostic marker for liver cancer survival. Out of 19219 gene expression 6032 gene expression was considered significant at $p < 0.05$. However these are raw p-value without taking into account Cox regression model was run for 19219 gene expression.

3.3 Storey and Tibshirani [2] based multiple hypothesis correction

3.3.1 Estimating π_0 estimates on p

Storey and Tibshirani [2] provided the mathematical implementation for false positive discovery rate. Using the implementation as discussed in the paper we calculated the π_0 estimates. Using the cubic

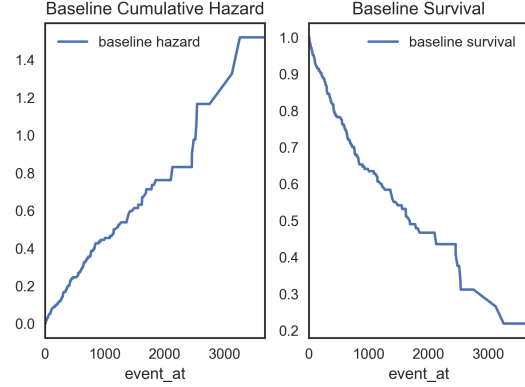


Figure 2: The baseline cumulative hazard function for a mock gene transcript *ENSG00000000003* The p-value for the transcript was 0.43 with hazard coefficient of -0.0057 indicating that it is not a reliable prognostic marker of liver caner prognosis.

fitting π_0 and λ values we calculated the $\hat{\pi}_0$ for the p values obtained from Cox model for each transcript as shown in figure 3. The $\hat{\pi}_0$ was estimated at 0.56.

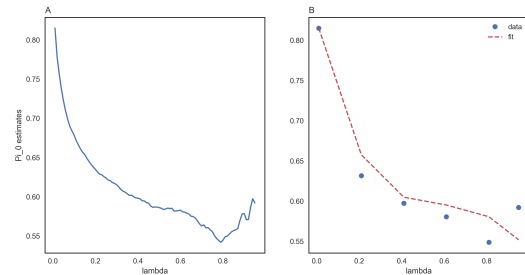


Figure 3: A) π_0 estimates of λ for 0.01 to 0.95 using the Storey and Tibshirani [2] implementation in python. B) Cubic spline fitting to π_0 and λ estimates. The blue points are subsets of real data and dotted lines are corresponding fitted cubic spline

3.3.2 Estimation of q-values based on $\hat{\pi}_0$ values

For each p-values of the transcript gene expression, a $\hat{\pi}_0 p_i$ was implemented in python in order to find

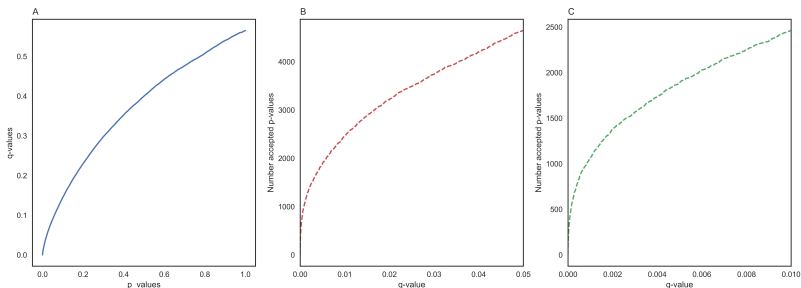


Figure 4: A) p and q values for different gene expression features. B) The number of p values accepted at q values 0.05 . C) The number of p values accepted at q values 0.01

the estimated q values for the i th significant features using the equation as defined in Storey and Tibshirani [2]. As shown in figure 4 we are able to control the false positive rates for p values calculated. We are able to find at least 4649 expressed gene transcript at $q < 0.05$. However considering the number of features tested we can further go to stricter cutoff such as $q < 0.01$ which would yield us 2464 gene transcript expression as prognostic marker of liver cancer survival.

4 Conclusion

In the current project we took a publicly available TCGA liver cancer expression and survival time to model a Cox’s proportional hazard model. At FDR 1% and 5% we found 2164 and 4649 gene transcripts as prognostic for survival of the patients. Interestingly, one of the top gene transcripts ENSG00000183520 is a prognostic marker for liver cancer as indicated into human protein atlas [3].

References

- [1] Cam Davidson-Pilon. Lifelines, 2014.
- [2] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.
- [3] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.