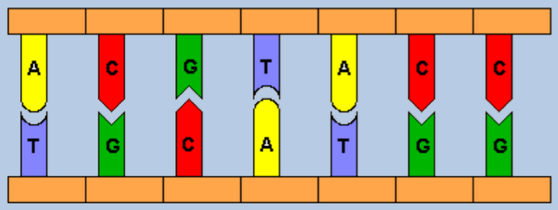
Machine Learning Approach for Clustering Lung Cancer Patients

Jozef Porubcin, Ting Jin, Daifeng Wang
Department of Biomedical Informatics
Rivermont Collegiate
Stony Brook University

Introduction

Gene Expression Levels

- Humans share 99% similar genes
- Gene expression levels differ significantly between humans, which is why we do not all look the same
- The dataset used contains more than 500 patients and over 20,000 genes
- The *k*-nearest-neighbors algorithm can be used to impute missing values



Feature Selection

PCA, RFE, and Correlation

- PCA (Principal Component Analysis) can reduce the number of features (in this case, 20,435) within a dataset to a number suitable for computation (15 PC's – Principal Components)
- RFE (Recursive Feature Elimination) ranks the features in a dataset by importance
- Correlation is a statistical method that can calculate whether/how strongly a pair of variables are related

Method

PAM

- PAM (Partitioning Around Medoids) is a method of separating data into *k* clusters (*k*, being an integer) based on the cluster with its nearest mean

The Two Pipelines

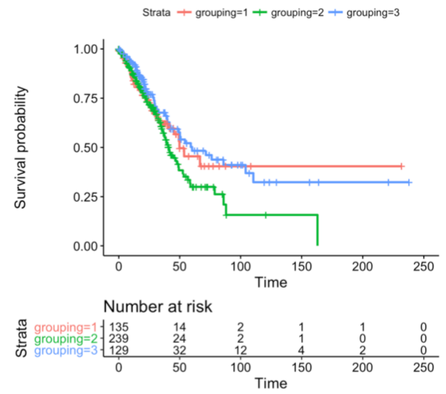
PCA • Selects features → RFE • Ranks PC's → PAM • Clusters patients → Survival plot

PCA • Selects features → Correlation • Calculates correlation between genes and PC's → PAM • Clusters patients → Survival plot

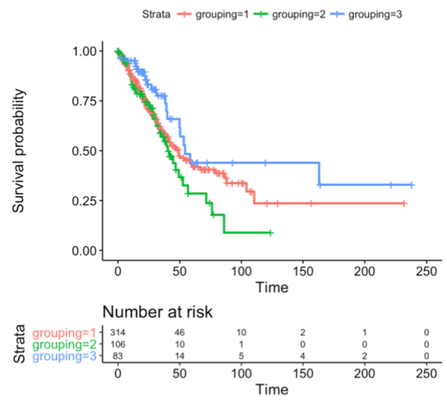
- Patients were clustered in two different ways; by top PC's and by top genes
- Clustering using PAM by top PC's involves RFE
- Clustering using PAM by top genes requires calculating the correlation between the genes and PC's
- The top 15 PC's and the top 150 genes were used to generate the first and second survival plots, respectively
- survfit() creates the model to be plotted
- ggsurvplot() draws both survival plots with information from the model

Results

Survival Plot with Top PC's



Survival Plot with Top Genes



Conclusion

- The differences between the clusters in both survival plots show that clustering patients by top PC's and top genes is effective

References

http://www.geneticsrus.org/DNA/dna2.php
https://ghr.nlm.nih.gov/primer/basics/dna
http://www.sthda.com/english/rpkgs/survminer/
https://cran.r-project.org/web/packages/survminer/vignettes/Informative_Survival_Plots.html
https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/cor.test