Term Project Proposal
High-Grade Prostate Cancer Prediction using DNA Methylation Data

Tanjung Dion (201899213)
Fawwaz Dzaky Zakiyal (201899213)

Bioinformatics (Fall 2018)

1 Introduction

Prostate cancer is a disease that affects a significant proportion of the male population, but in most cases, the disease is harmless. However, for the small portion of the population with high-grade prostate cancer, the disease can be extremely debilitating, causing painful symptoms and even death. Prior studies have been performed, showing how hereditary factors, alcohol consumption, sexual activity, family history and race can each play a role in developing high-grade PCa. Over the past years rapid advances in next-generation sequencing technology has led to the timely advent of The Cancer Genome Atlas (TCGA) project which provides the most comprehensive genomic data for various kinds of cancer. In the previous research indicates classification of patient samples using TCGA Exon expression or SNP datasets. However, recent study show that DNA methylation act as better bio-markers and help in improving the cancer prognosis [1].

2 Vocabulary

- **DNA methylation**: the process of adding methyl groups to DNA, in this process modification of covalent nucleotides in the human genome, namely cytosine and also guanine. It is one form of epigenetic modification which take an important role in the development of cancer.
- **Gleason Score**: a measure (from 2-10) of the aggression of prostate cancer cells based on clinical pathology of prostate tissue.
- **High-Grade Prostate Cancer (HGPCa)**: a form of prostate cancer that is likely to metastasize and generally results in poor patient outcomes mortality, complications, and long-term disease free survival). A gleason score from 8-10 is indicative of High-Grade prostate cancer, while scores below 8 are considered not as severe.

3 Methodology

3.1 Dataset

The data used in this paper is publicly available through the National Cancer Institute GDC Data Portal, and can be found online under Project TCGA-PRAD. This data set consists of genomic information and clinical pathology reports belonging to roughly 500 patients who have been diagnosed with prostate cancer (a Gleason Score anywhere between 2 to 10). It contains of both clinical (recurrence, survival & treatment resistance) and molecular profiles (Exon (mRNA) expression, DNA methylation, Copy Number Variations (CNV) & Single Nucleotide Polymorphism (SNP)) for both tumor samples and normal controls. Data from TCGA has improved the ability of oncologists to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease. The clinical pathology reports were also critical for our analysis because they contain the Gleason scores for each patient, which allows us to determine whether a patient has HGPCa [2]. The dataset detail can be seen at Figure 1.

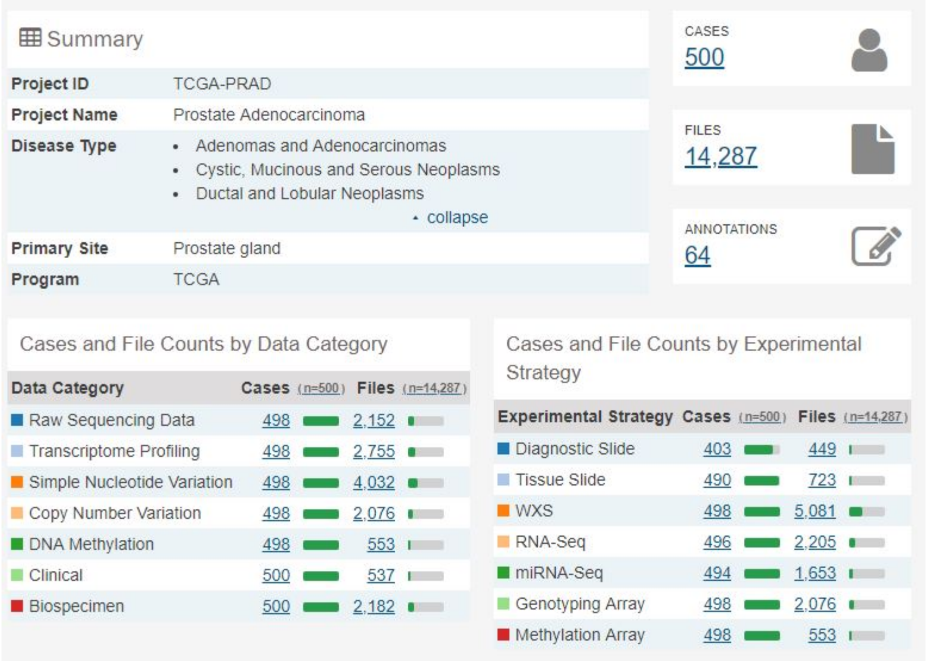
3.2 System Design

The process consist of data preprocessing of TCGA-PRAD dataset, and continue with feature selection of extracted DNA methylation data. After that, training the classifier models, testing and evaluating those models. The system design can be seen at Figure 2.

3.3 Preprocessing

Significant amounts of pre-processing needed to be performed in our to make the data usable in any way. First, we had to consolidate our data across hundreds of clinical pathology reports and DNA methylation profiles. After consolidation, we constructed a feature matrix in which each row corresponds to a patient and each column corresponds to a DNA methylation.

We will also applied the adjustments described below to each column in order to normalize our data and using Principal Component Analysis (PCA) as feature selection algorithm to eliminate the useless DNA methylation and reduce the search space.



Summary

Project ID	TCGA-PRAD
Project Name	Prostate Adenocarcinoma
Disease Type	• Adenomas and Adenocarcinomas • Cystic, Mucinous and Serous Neoplasms • Ductal and Lobular Neoplasms ▴ collapse
Primary Site	Prostate gland
Program	TCGA

CASES 500

FILES 14,287

ANNOTATIONS 64

Cases and File Counts by Data Category

Data Category	Cases (n=500)	Files (n=14,287)
■ Raw Sequencing Data	498	2,152
■ Transcriptome Profiling	498	2,755
■ Simple Nucleotide Variation	498	4,032
■ Copy Number Variation	498	2,076
■ DNA Methylation	498	553
■ Clinical	500	537
■ Biospecimen	500	2,182

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=500)	Files (n=14,287)
■ Diagnostic Slide	403	449
■ Tissue Slide	490	723
■ WXS	498	5,081
■ RNA-Seq	496	2,205
■ miRNA-Seq	494	1,653
■ Genotyping Array	498	2,076
■ Methylation Array	498	553

Figure 1: Dataset Detail

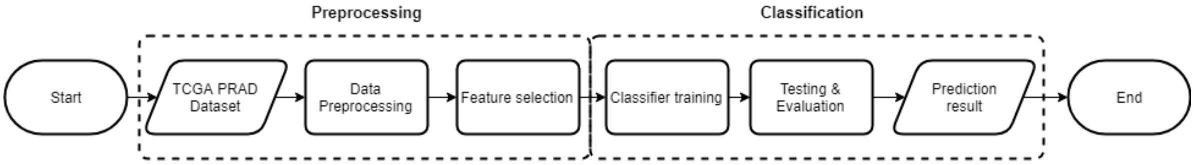


Figure 2: System Design

3.4 Classifier Algorithms

Perform a classification for cancer prediction :

1. K-Nearest Neighbor (KNN)
2. Support Vector Machine (SVM)
3. Multilayer Perceptron (MLP)

3.5 Testing & Evaluation

This project will finding a model that has the maximum classification accuracy, given by the Equation 1, where y_i' is the predicted label for instance i, y is the true label for instance i. I is an indicator function, if the statement is true, it will return 1, and vice-versa.

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i' = y_i) \tag{1}$$

References

[1] Hao, Xiaoke, et al. *DNA methylation markers for diagnosis and prognosis of common cancers.* Proceedings of the National Academy of Sciences 114.28 (2017): 7414-7419.

[2] GDC. Portal.Gdc.Cancer.Gov, 2018, *https://portal.gdc.cancer.gov/projects/TCGA-PRAD.* Accessed 2 Nov 2018.

[3] Ang, Jun Chin, et al. *Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection.* IEEE-ACM transactions on computational biology and bioinformatics 13.5 (2016): 971-989.