Project Proposal

Yun Mai

July 15, 2017

Project Proposal

Introduction

Cancer encompasses a broad spectrum of diseases (>100) that arise from somatically acquired genetic, epigenetic, transcriptomic, and proteomic alterations that have accumulated in the genomes of cancer cells. These alterations are implicated in hallmark oncogenic cellular processes that are characterized by, e.g., sustained proliferative signaling, resistance to apoptosis, induction of invasion and metastasis, and neoangiogenesis . The somatic loss-of-function or gain-of-function alterations are overrepresented in specific genomic regions, which could indicate their potential suppressive or oncogenic roles, respectively. However, it must be noted that somatic mutations occur on different genetic backgrounds and can sometimes interact with germline mutations, which could modify predisposition to cancer when such mutations occur in cancer-associated genes.

Recent advances in technologies for high-throughput genome analysis, such as microarray-based methods and next-generation sequencing (NGS), have enhanced progress in the field of oncogenomics . These tools were fundamental for the initiation and development of multi-centered cancer genomic projects, such as (i) the Wellcome Trust Sanger Institute's Cancer Genome Project (CGP) , (ii) The Cancer Genome Atlas (TCGA) , and (iii) the International Cancer Genome Consortium (ICGC) cancer genome projects. These projects have been launched for genome-wide analyses of genetic, epigenetic, transcriptomic, and proteomic alterations in hundreds or even thousands of cancer samples. Their general aim is to provide publicly available oncogenomic datasets for the better understanding of the molecular mechanisms that underlie cancer and for the assessment of the influence of specific alterations on clinical phenotypes. Application of the appropriate pipeline for computational interpretation and thought-provoking visualization of the results of oncogenomic projects is crucial to exploring the multidimensional character of genome-wide cancer data. In response to this need, a number of oncogenomic portals were created to assist with accessing the abundant cancer datasets. These portals gather and facilitate the analysis of data with regard to small-size mutation, copy number variation (CNV), methylation, and gene/protein expression. Moreover, they offer a wide range of analysis tools that include the testing of correlations of specific genomic alterations with available clinical information.

I will explore the ways to illustrate the relationship between gene mutation and cancer inorder to see how visualization can facilitate scientists from different cancer-associated fields, including molecular and clinical oncologists, epidemiologists, and bioinformaticians, with the extraction of meaningful information from expanding oncogenomic sources.

The Data

This project uses the public available genomwide sequencing data from cBioPortal (http://www.cbioportal.org/data_sets.jsp) and COSMIC(http://cancer.sanger.ac.uk/cosmic/download) A variety of data regarding cancer types and subtypes, oncogenic molecular pathways and cancer-associated genes of interest will be used.

The Tasks

For this project, we have chosen to focus on two tasks:

Task 1: Show the basic information about the gene and the detailed analysis of the gene alterations in histogram and a table.

Task 2: Show the 20 most frequently mutated genes. Show the CNV plotto summarize the copy number variations across the whole genome of the chosen cancer type. Show the Mutation Matrix presenting alterations in the most frequently mutated genes in the tumor samples that have the highest number of alterations.

Task 3: Circular plot of all of the alterations (coding mutations, gene expression and CNV) that are detected in an individual exemplary sample (TCGA-A6-5657-01) of the cancer.

For this visualization we will be utilizing a suite of software applications for the purpose of Processing, Analyzing, and finally Visualizing these data.

Programming Language:

For this task we have settled on Python for its easy to read syntax, and wide range of available tools and frameworks.

Visualization Frameworks:

We will visualize the final data using a variety of visualization tools and frameworks.

matplotlib

Plotly

Seaborn

cufflinks

bokeh

Dash

Statistical Analysis Packages:

We may use the neuroimaging statistical analysis package statsmodels in order to process the data.

Expected Results and Evaluation

I will have to download the data from the provided source. Data cleaning will consist of extracting the relevant attributes of interest and applying.

Almost without question I will have to convert the data into various formats in order to take full advantage of the tools we have chosen to visualize our application.

Following those procedures, I will implement the described visualization.