Rheinisch Westflische Technische Hochschule Aachen
Informatik 5 (Information Systems)

Master Thesis Proposal

Capturing Data Curation Process in Cancer Genomics

Jasim Waheed Ansari

Aufgabenstellung: Prof. Dr. Stefan Decker
Betreuer: Dr. Oya Deniz Beyan

Aachen, den May 27, 2017

Contents

Chapter 1

Introduction

Cancer is one of the leading causes of death around the globe. It is considered to be the most complex disease that human species has to deal with. It is characterized by undesirable growth in cells which are caused by changes in genomes and exposure to environment[2]. Each cancer type has unique architecture of genetic variation - somatic mutations, copy number alterations, gene expression profiles and different epigenetic aberration. Therefore, there is a strong need for better diagnosis, personalized medicines for patients have arose, which correlates with better understanding the changes at genetic level in tumors. All thanks to technological advancements in high-throughput techniques for genome-wide interrogations such as Microarray and Next-Generation Sequecing(NGS) platforms have provided means to target personalized healthcare solutions for patients[5] [4].

1.1 Data Integration Challenges in the era of Cancer Genomics

With the deluge of multiple projects from different organizations catering to Cancer Genomic has grown over years, one such is The Cancer Genome Atlas(TCGA). It comprises of genomic, epigenomic, and proteomic data from over 10 thousand samples from 33 kinds of cancer, with a target to support bioinformaticians and medical researchers in understanding molecular foundation of cancer development. Since the data available in TCGA qualifies the 3Vs i.e. Volume, Velocity and Veracity principle of Big Data, integration of comprehensively characterizing the whole genomes at low cost is a big challenge[1][3]. Performing mining on the big data extracted through TCGA is overwhelmingly difficult by the traditional tools like R and Python. Furthermore, the data collected in TCGA site exists in native raw form, therefore there is a need to perform further preprocessing which needs background in both biological domains as well as technical expertise for programming. TCGA provides ample of intuitive web-based tools, such as TCGA-Assembler, cBioPortal, Firebrowse etc which supports clinicians and researcher in making knowledgeable insights of all the data types in a significant manner. Visual analytical tools has features which offers integrative data mining capabilities in the following modes:

- **Discovery mode:** Finds global correlations across different data types.
- **Confirmation mode:** Analysis on results from few genes.

Visualization tools such as cBioPortal, Firebrowse often offer one of the two modes of study, such as:
a) cBioportal features study of multiple studies at the same time
b) FireBrowse offers gene level study
Both of them lack the ability to incorporate into a single analysis. Whereas, TCGA-Assembler offers no visual interface and requires scripting knowledge to deal with the datasets for analytical purposes.
Therefore, through our work, we develop a scalable curated-data analytics pipeline infrastructure built with visualization dashboard on top for analytical purposes, commonly known as Data Lake, with an ultimate aim for comprehensive analyzing and visualizing the most common data types provided by TCGA.

1.2 Thesis Goals & Outcome

Specific aims of my thesis are as follows:

- build a scalable centralized repository, ingesting TCGA analytical data sources such as cBioPortal, Firebrowse
- develop a data management solution for this infrastructure which:
  i) captures the metadata at all the levels i.e. business, technical and operational metadata
  ii) offering capability to data profiling, reporting and governance
- extract value from this platform by facilitating data visualization with interactive dashboard

Chapter 2

Background & Related work

2.1 What is Cancer Genome?

2.2 Current state-of-art

2.2.1 The Cancer Genome Atlas

cBioPortal

Firebrowse

2.3 Big Data Fundamentals

2.3.1 Technologies for Big Data

2.3.2 Shortcomings in traditional Data Warehousing

2.3.3 Need for Data Lakes in Big Data

2.3.4 Big Data Visualization

Chapter 3

Solution

3.1 Development of Big Data Lake Platform using Kylo

3.1.1 Data Ingestion Pipeline and Design with Apache Nifi

3.1.2 Data Standardization and Validation with Apache Spark

3.1.3 Metadata Tracking

3.2 Development of interactive dashboard using Apache Zeppelin

Chapter 4

Evaluation

4.1 Distinguishing Breast Cancer Tumor Subclasses of Gene Expression Patterns with Clinical Implications

4.1.1 Background

4.1.2 Dataset in Detail

4.1.3 Identification of Tumor Subtypes by Hierarchical Clustering

4.1.4 Correlation to Clinical Results

4.2 Results & Discussion

Chapter 5

Time Plan

Bibliography

[1] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[2] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, mar 2011.

[3] Ju-Seog Lee. Exploring cancer genomic data from the cancer genome atlas project. *BMB reports*, 49(11):607, 2016.

[4] Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[5] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.