Rheinisch Westflische Technische Hochschule Aachen
Informatik 5 (Information Systems)

Master Thesis Proposal

Semantic Data Profiling in Data Lake for Cancer Genome

Jasim Waheed Ansari

Supervisor: Prof. Dr. Stefan Decker
Advisors: Dr. Oya Deniz Beyan, Dr. Michael Cochez, Naila Karim

Aachen, den July 8, 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[7]. Each cancer type has unique architecture of genetic variation - somatic mutations, copy number alterations, gene expression profiles and different epigenetic aberration. Therefore, a strong need for better diagnosis, personalized medicines for patients have arose, which correlates with better understanding of 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 wealth of multi-disciplinary, multi-institutional pilot projects in cancer studies[14] [9].

Various unique projects catering to cancer genome study such as 1000 Genome Project[3][4], Encyclopedia of DNA Elements Project (ENCODE)[5], Immunological Genome Project (ImmGen)[13], The Cancer Genome Atlas (TCGA)[8] which aims at investigating biological systems at several levels through the creation of humongous datasets with different data types. These projects motivates the bioinformaticians such as development of novel data integration methodologies to explore hidden insights.

With respect to current state-of-the-art for data integration platforms, data lake promises to addresses the onslaught of Big Data generated across the projects[6]. However, one of the biggest challenges in data lake is to prevent turning into a data swamp or silos. Which means, the lake turns into a set of disconnected data pools because of lack of semblance of information governance. Presently, 70% of the time spent in data analytics projects is into informational discovery of the data - identification, location, integration and re-engineering[15]. Moreover, especially for cancer genomic data, where the users must have a first hand knowledge in biomedical area since its metadata alone cannot provide the full meaning out of respective datasets. There are dearth of novel methodologies that can point out, for example - genes with name "TP53" and "BCC7" or disease with name "Breast Cancer" and "Breast Tumor" means the same. Additionally, in context of conceptual hierarchy, the patient's clinical data set has a broader term as a Person or narrower terms such as Outgoing, Incoming or Doner. Thus, the use of knowledge organization system from both biomedical domain as well as generic is needed to understand the data more unambiguously. This would assist the users to find, explore, understand, and trust

the data at ease. Such domain knowledge and thesauri are captured as specification of conceptualization(or commonly known as ontologies) in BioPortal[11] and Simple Knowledge Organization Systems(SKOS)[10] etc.

Keeping in consideration of our focus of work in relation to semantic discovery and management of metadata around the informational content, we focus on profiling aspect. But, existing profiling capabilities revolves around catching statistical summaries of anomalies and does not offer semantic aspect. This motivates us to an idea of enabling semantic profiling in data lakes.

Semantic Profiling is a methodology that exploits semantic-based tools and ontologies in order to derive lucid understanding of the information being stored in current systems. This approach is much more rigorous than former traditional profiling techniques for the following reasons:
i) Unlike traditional profiling which captures as much anomalies as you can until you stop, whereas, in semantic profiling once you select the domain of study, you stick by the rules and guidelines.
ii) If a user perform semantic profiling on one system and would like to combine the results of first system with those from other system, it can easily be facilitated through its feature of reusability.
iii) Semantic profiling allows setting up of testing hypothesis to monitor the data lake in production scenario which can detect semantic drift.

Therefore, through this thesis work, we extend current profiling efforts in the data lake solutions in a much more rigorous way i.e. through semantical enrichment using ontologies in Figure 1.1. This would open the door in answering questions about data sources like:

- Does the biological data adhere to particular standards such as gene and disease ontologies?
- Is there a way to capture controlled vocabularies, taxonomies and thesauri for clinical information of patients?

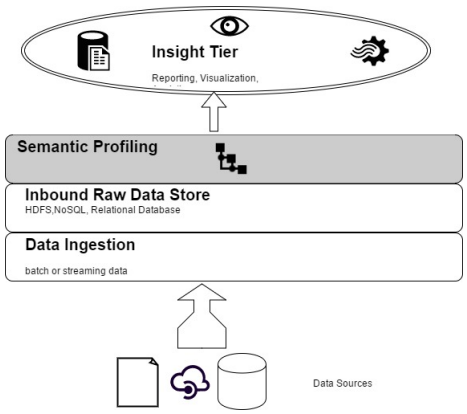

Figure 1.1: Need to extend the current semnantic tendencies of data lake

1.1 Thesis Goals & Outcome

Specific aims of my thesis are as follows:

- Extend and develop current profiling efforts in our state-of-the-art data lake with:

  a) a systematic process of annotating the ingested data's schema[1]

  b) a systematic approach of extracting, managing and exploiting metadata of the datasets' information with ontologies(for example, BioPortal, SKOS) through ontology alignment techniques[12]

- Analysis of the results

  a) Develop a dashboard containing provenance to visually debug the quality of resulting data

Chapter 2

Background & Literature Survey

In this chapter, we look into basic background on cancer genome and contemporary related work.

2.1 What is Cancer Genome?

2.2 State-of-the-art cancer genome repositories

2.2.1 The Cancer Genome Atlas(TCGA)

2.3 Knowledge Organization Systems

2.3.1 Simple Knowlege Organization Systems (SKOS)

2.3.2 Bioportal

2.4 Related Work

2.4.1 The DQM Tool by Talend

2.4.2 Data Lake Content Management

2.4.3 Kylo: Data Lake solution

Chapter 3

Solution

3.1 Extend and develop current profiling efforts in Kylo, a data lake solution

3.1.1 Systematic process of annotating the ingested data's schema[1]

3.1.2 Systematic approach of extracting, managing and exploiting metadata of the datasets' information with ontologies

3.2 Analysis of the results

In order to make claims about our developed methodology is fool-proof, we need to analyze the results through work-flows In this work, we utilize Provenance as an infrastructure for debugging and consistency check on our results.

3.2.1 Develop a dashboard containing provenance to visually to visually debug the quality of resulting data

We perform in-depth analysis of the semantically annotated result through visualization in a dashboard, taking into two aspects in consideration:

a) Tracking our current results against the newly scheduled run on the data source
b) Tracking if the process is identical from two different data stewards

Chapter 4

Evaluation

This section is to guarantee that the results we achieved through our work is not random result but meeting our thesis objectives.

The evaluation will be carried out using the Broad GDAC Firehose dataset into two separate categories[2]:
a) molecular datasets
b) clinical datasets

It is to be noted that the stability of the results is very important. Hence, we will test the feasibility of our work against previous methodologies on the same dataset.

Chapter 5

Time Plan

Bibliography

[1] Philip A Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011.

[2] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized tcga data from broad gdac firehose 2016_01_28 run, 2016.

[3] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[4] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[5] Joseph R Ecker, Wendy A Bickmore, Inês Barroso, Jonathan K Pritchard, Yoav Gilad, and Eran Segal. Genomics: Encode explained. *Nature*, 489(7414):52–55, 2012.

[6] Huang Fang. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on*, pages 820–824. IEEE, 2015.

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

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

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

[10] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. Skos core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, pages 3–10, 2005.

[11] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

[12] Marcos Martínez Romero, José Manuel Vázquez Naya, Javier Pereira Loureiro, and Norberto Ezquerra. Ontology alignment techniques. In *Encyclopedia of Artificial Intelligence*, pages 1290–1295. IGI Global, 2009.

[13] Tal Shay and Joonsoo Kang. Immunological genome project and systems immunology. *Trends in immunology*, 34(12):602–609, 2013.

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

[15] Ignacio G Terrizzano, Peter M Schwarz, Mary Roth, and John E Colino. Data wrangling: The challenging yourney from the wild to the lake. In *CIDR*, 2015.