Michael Lee, Jonathan Wolf

CSE 446

Project Proposal

**Project title**- “Predicting Gene Expression Levels across Cancer Afflicted Patients”

**Data set**- The Cancer Genome Atlas (TCGA).

Our data is stored in a Google BigQuerry Table that we can access easily via SQL. This data set was provided by the Institute for System Biology and contains information for thousands of tumor samples. Specifically, it contains relevant biological information for each sample such that we can pursue our question.

­

**Project Idea**:

With this data set we plan to predict gene expression for cancer afflicted patients. Gene expression refers to the process by which a functional gene (ie. insulin,lactose, etc.) is produced by the information from a gene (that being a piece of our DNA). This process is ubiquitous in all organisms and understanding it better is all the more pertinent in a cancer data set.

Gene expression involves many levels of regulation which we seek to use as features in our project. Some example features include DNA methylation, miRNA expression levels, patient data such as age/weight/race. We plan to examine a known gene pathway for a specific cancer type. Meaning that we plan to predict gene expression for a cluster of related genes that when mutated are known to induce cancer.

**Software you will need to write.**

To attain the data, we will need to use SQL to process it from the BigQuerry table. I have received many tutorials from correspondence at the ISB for working with BigQuerry and furthermore, my teammate is competent with SQL and this should not pose a problem.

We are considering two approaches to the software we need to write for the project. One approach is to download the data that we need using SQL and then program our ML scripts in python. Another approach we are considering is writing our scripts in SQL to run on BigQuerry. Given its analytical power it is something that we are considering though, the first approach may me be more feasible given the time constraint.

**Papers to read.**

<http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r53>

<https://cancergenome.nih.gov/publications>

**Teammate**:

I will be completing the project with Johnathan Wolf.

**Milestone**: What will you complete by the milestone? Experimental results of some kind are expected here.

We plan to run at least one or two ML algorithms on the data and get confirmation for the underlying principle of the project.

Specifics:

Cancer Type(s): We will take multiple cohorts and combine them into one

Features:

miRNA expression

DNA Methylation

Clinical Data:

Age

Smoking-Status

Race

etc..

Mutations In a specific pathway.

\*would need a list of known cancerous proteins

\*would need a list of genes that are contributors.

We will predict gene expression across **all** genes, we can consider a random subset of genes of the 20,000, gather RNA-expression levels for those patients:

Example: Gene X person Y

Label: mRNA expression level gene X in person Y

Features:

Proximity of nearest mutation **|** DNA-methylation

age | mRNA expression of other proximal genes |