Michael Lee

CSE 446

Project Proposal

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

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

­

**Project idea**:

Predict expression level of genes across cancer afflicted patients

I would think that it would be best to look at a specific cancer and genes in a signaling pathway.

Features of interest:

Types of mutations involved- SNP, DEL, INS

\*TATA box mutations

Transcription Start Site Mutations

Histone Modifications

CpG Information

\*PolyA Mutations

miRNA expression levels

TF motif mutations

Patient Data- gender, age, ethnicity, days until deceased, other factors

Other ideas to include:

-There is a reference genome in the data set provided to which we can compare, ie. TCGA data and reference data could be analyzed side-by-side.

- We could also learn on the cancer patients biological characteristics and analyze the reference genome and make predictions on that reference genome’s susceptibility for cancer.

-Another analysis we could perform is PCA on the two studies and see which factors seem more powerful in one study versus the other.

**Software you will need to write.**

ML Algorithm in google BigQuerry.

**Papers to read.** Include 1-3 relevant papers. If you are doing something different then one of the suggested projects, you will probably want to read at least one of them before submitting your proposal.

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

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

**Teammate**: will you have a teammate? If so, whom? Maximum team size is two students. One proposal per team.

I will be completing the project with Johnathan,

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

We expect to have run an ML algorithm on our data set for at least one gene on the cancer afflicted patientss.