03-712 Project Proposal

Team Members: Frank Lin, John-Krill Burger, Prateek Tandon, Vineet Joshi

Introduction:

A tumor is an abnormal growth of cells. A benign tumor does not have the ability to invade neighboring tissue while a cancerous tumor does. A tumor is classified as metastatic when it has spread from its organ of origin to other organs. Cancers originating in different organs have drastically different disease progressions, prognosis, and response to treatment. Accurate diagnosis of the tumor type is critical to determining the best course of treatment and is usually accomplished by removing a small piece of the tumor by fine-needle aspiration and then using staining, cytochemical analysis, fluorescence in situ hybridization (FISH), and other molecular tests to determine the type and stage of the tumor. While each individual cancer subtype has its own set of features that pathologists look for, the feature that most drastically changes the course of treatment for all tumors is whether it is metastatic or not. Therefore, the use of molecular markers for preoperative diagnosis of metastatic tumors is an important area of research that could have immediate clinical impact. In this study, we investigate the use whole genome and exome sequencing data from human primary tumors to identify metastatic mutations and use them to build a classifier for metastatic tumor prediction.

Aims:

- To develop a classifier that is capable of dividing tumor exome sequences into the categories of metastatic and non-metastatic tumors.
- To perform cross validation of the classifier and improving the quality by sensitivity vs specificity of the output.
- To perform n to 1 (where n >= 1 and represents the number of mutations considered) correlation analysis followed by statistical significance testing between mutations identified to be belonging in the metastatic class data with an aim to identify to solve an instance of the Hard discrete optimization problem - finding the minimum test set of the prognostic markers for the metastatic tumor diagnostic purposes.

Work Plan:

- **Week 1:** Obtain and organize data from University of Pittsburgh's department of pathology as well as supplementary data from the TCGA cancer genome atlas. The data consists of pairs of metastatic tumors and non-metastatic tumors sequences from patients with breast, prostate, thyroid, renal, esophageal, bladder, and pancreatic cancers.
- **Week 2:** Filter data by removing noise followed by running alignments and SNP identifications.
- **Week 3:** Algorithm decision for the classifier. Reasoning with the correctness of the algorithm.
- Implementing the algorithm for the classifier.
- **Week 4**: Inputs and outputs outline, choice of data structures to be used. Data parsing into appropriate data structures. Implementing continued and integration with the parser.
- **Week 5:** Running the algorithm on the data we have. Perform correlational analysis and significance testing.
- **Week 6:** Consolidate findings, generate conclusions and writing reports, preparing presentation.

Team Member Contribution: Each member will be assigned a different cancer type and he has to build a parser according to the data available.

- The statistical significance of the output will be assigned to Vineet.
- The sequencing data acquisition and inference will be done by John.
- The noise reduction and SNP identification and correctness of the algorithm analysis will be done by Frank
- The classifier building will be in general by contribution from all but Prateek will be responsible for maintaining it and consolidating individual parts developed by other team members.