Assignment 3 - Project Proposal

Arun Ram S (arunsank)

March 3, 2016

1) PROJECT TITLE - PREDICTION OF CANCER CAUSING SOMATIC MUTATIONS IN GENE SEQUENCES

2) NAMES

i) Arun Ram S - arunsank@umail.iu.edu
ii) Mithilesh Nanjamanaidu SR - minsrini@indiana.edu
iii) Amol Bhagwat - arbhagwa@indiana.edu

3) OBJECTIVES AND SIGNIFICANCE

i) Goal

The goal of the project is to impart predictive modeling by employing prediction algorithms like random forest on the COSMIC (Catalogue of Somatic Mutations in Cancer) dataset. COSMIC is currently the most comprehensive global resource for information on somatic mutations in human cancer, combining curation of the scientific literature with tumor, resequencing data from the Cancer Genome Project at the Sanger Institute, U.K. Almost 4800 genes and 250000 tumors have been examined, resulting in over 50000 mutations available for investigation (Forbes, S.a., G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J.w. Teague, P.a. Futreal, and M.r. Stratton., 2008). We would use a neutral dataset to distinguish the somatic mutations (driver mutations) from the other mutations (passenger mutations), which in most situations impose no harmful effects to the human body.

ii) Importance

This kind of predictive modelling is very important because with the advancement in technology the amount of missense mutations being detected is increasing at an alarming rate. Amongst the missense mutations being identified, there may be many amino acid substitutions that are not at all harmful and pose no effect on the human body. Hence there arises a need for a medium to distinguish harmful somatic mutations from the other harmless mutations that may occur in the cells.

iii) Motivation

We were inspired by the advancement in the study of the process of acquired mutations and the knowledge that cancers can be predicted using data mining techniques triggered us to start this project. Acquired mutations which involve somatic cell division may sometimes be the cause of cancer causing agents. Although there can be inherited components which may lead to cancer promoting acquired mutations, we are mainly going to concentrate on predicting the somatic mutations.

4) BACKGROUND

The human body consists of somatic and germline cells. Mutations accumulate in both these cells, while only the germline mutations are inherited in humans. Somatic cells as seen in the embryo may contribute in formation of almost all the human body parts. A somatic mutation is a mutation in a somatic cell in contrast to a mutation in a germ cell. Somatic mutations are caused by somatic cell divisions. These kind of mutations are not passed on to the next generation unless they are cloned. Somatic mutations are found to affect cells in different ways than in which they affect the organisms. For instance, a somatic mutation which may be harmful to the cell may be of no impact to the organism. On the other hand, somatic mutations which may be beneficial to the cells may be a major disease causing agent to the organism. Recent research has confirmed that there are several somatic mutations that may cause diseases like cancer in the organisms.

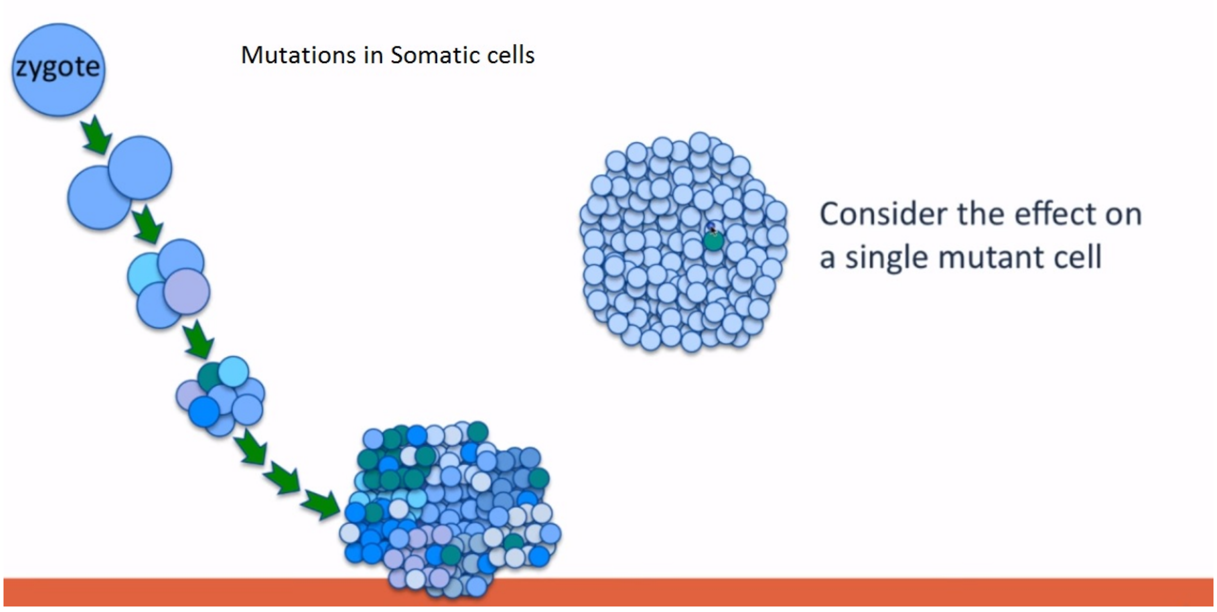

Figure 1: Depiction of mutations in Somatic cells

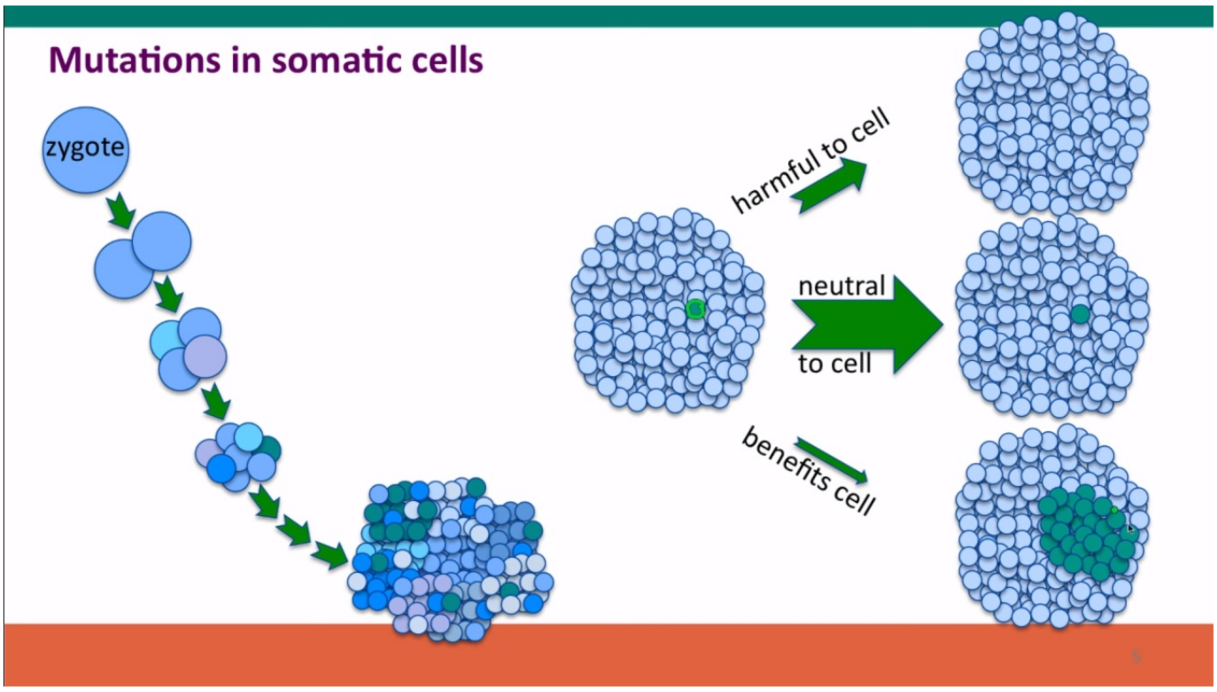

Figure 2: In chronological order, cells that are harmless, neutral and harmful to the organism

As we would know the body tissues are a product of somatic mutations. Hence in a human body, there are always three important effects of a somatic mutation. There may be a mutation that may be harmful to the cell but less harmful to the body. There may be neutral mutations which do not pose any threat at all to the body. Beneficial mutations which may be beneficial to the growth of the cells but may deter the health of the body. One such example is the cancer tumors which are a result of rapid growth of tumor cells within the organism.

5(A) DATA

We plan on using datasets describing neutral and somatic mutations in nucleotide sequences. We have in our possession over 55k samples of neutral mutations which was obtained from the Single Nucleotide Polymorphism (SNP) database. The data is in FASTA format. The datasets describing somatic mutations will be obtained from COSMIC (Catalog of Somatic Mutations in Cancer) or TCGA (The Cancer Genome Atlas). The data consists of a mutation identifier, nucleotide sequence and a mutation description which consists of a position, the present amino acid, and a replacement.

A snapshot of the data as viewed in SnapGene Viewer is shown below -

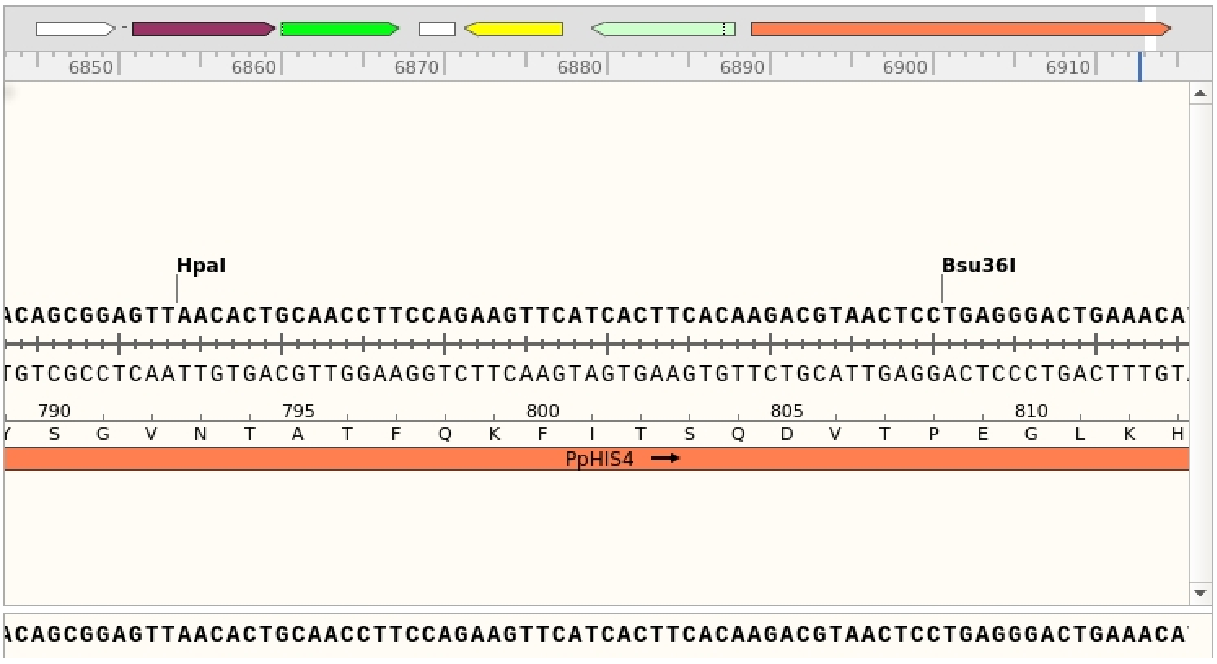
6850
6860
6870
6880
6890
6900
6910
HpaI
Bsu36I
ACAGCGGAGTTAACACTGCAACCTTCCAGAAGTTCATCACTTCACAAGACGTAACTCCTGAGGGACTGAAACA
TGTCGCCTCAATTGTGACGTTGGAAGGTCTTCAAGTAGTGAAGTGTTCTGCATTGAGGACTCCCTGACTTTGT
790
795
800
805
810
Y S G V N T A T F Q K F I T S Q D V T P E G L K H
PpHIS4
ACAGCGGAGTTAACACTGCAACCTTCCAGAAGTTCATCACTTCACAAGACGTAACTCCTGAGGGACTGAAACA

Figure 3: Flow Diagram

(B) i) PROPOSED METHODOLOGY

The methodology of the project is given in the flow diagram below:

LITERATURE

A thorough literature study will be performed by the team to get started with bio-informatics. Since the team is not experienced in the field of bio-informatics, a detailed study on the same will be done in order to know the basics of genome study, structure of genome data, types of genetic disorders,etc. Various links and articles listed in the reference section will be studied in detail.

DATA PRE-PROCESSING

Our scope of the project is to build a model which would classify a gene as neutral or somatically mutated. Separate sequence data are obtained from COSMIC/TCGA and Sin-

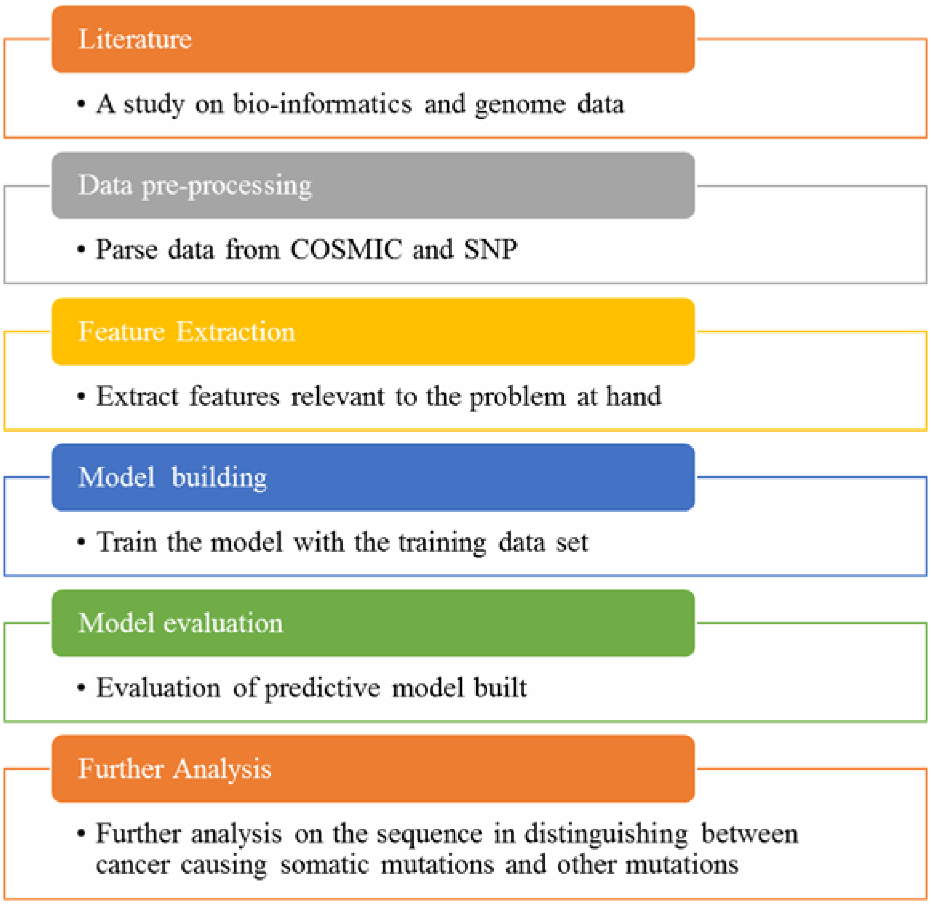
Literature
• A study on bio-informatics and genome data
Data pre-processing
• Parse data from COSMIC and SNP
Feature Extraction
• Extract features relevant to the problem at hand
Model building
• Train the model with the training data set
Model evaluation
• Evaluation of predictive model built
Further Analysis
• Further analysis on the sequence in distinguishing between cancer causing somatic mutations and other mutations

Figure 4: Flow Diagram

gular Nucleotide Polymorphism (SNP) databases (as discussed in section a) which contain somatically altered genes and neutral genes respectively with which we will train the predictors. Parsing the datasets is the first step. The genome data which consists of sequence of tokens are analyzed and their structure is determined. With this information, a parser builds a data structure such that a which is programming friendly. Both COSMIC DATA and SNP are parsed to be ready for modeling.

FEATURE EXTRACTION

Once the data set is available, features within our region of interest are extracted to train the dataset. These features are usually combinations of different sequences giving rise to a particular amino acid/nucleotide. Feature extraction can either be done using tools which automatically derive features from the sequences or by designing databases from which required sequences can be queried. Given the timelines of this project, we have decided to go ahead with feature extraction tools.

MODEL BUILDING

Once the features are extracted/selected, The examples are used to build a model. We have decided to go ahead with a Random Forest classification algorithm since it can be implemented on non-linear data, robust to noise in the data and also not computationally complex (though this depends on data, on comparing with bagging and boosting ensemble methods, it runs faster on the same datasets). Random Forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set. They also have the tendency to perform well with a large set of examples.

MODEL EVALUATION

The model will be evaluated by a common method of 10-fold cross validation and evaluation metrics such as precision, recall, f measure and ROC curve will be calculated to evaluate against other algorithms that were found in the literature.

FURTHER ANALYSIS

This methodology will give us the prediction of somatic/neutral classes and this could be extrapolated to predicting cancer causing somatic mutations.

ii)IMPLEMENTATION We are planning to implement this using either MATLAB or Python for the model building and evaluation, software/tools for parsing and feature extraction.

(C) EVALUATION

We propose running a 10-fold cross validation on the devised model. In 10-fold cross validation, the data is divided into 10 equally sized subsets. Each subset consists of samples such that there is an equal probability of all target classes. Out of these 10 subsets, 9 are used to train the model, and the evaluation is done based on the results obtained from the 10th subset. This process is done a total of 10 times, using a different testing subset each time. The performance metrics that we wish showcase are, accuracy, precision, recall, F measure and ROC curve.

(D) i) EXPECTED OUTCOME

A class predictor model which can accurately predict the class of the sequence as somatic/neutral which can lead us to cancer prediction model with an accuracy comparable to other classification algorithms already explored. If this fails, we have another idea - image segmentation by clustering aiding autonomous navigation.

(6) INDIVIDUAL TASKS

Arun - Design and Programming
Mithilesh - Research, Analysis and Programming
Amol - Programming and testing

(7) REFERENCES

i) Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David

Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." Nat Biotechnol Nature Biotechnology 31.3 (2013): 213-19. Web.

ii) Forbes, S.a., G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J.w. Teague, P.a. Futreal, and M.r. Stratton. "The Catalogue of Somatic Mutations in Cancer (COSMIC)." Current Protocols in Human Genetics
(2008). Web.

iii) Meneely, Philip Mark., and Philip Mark. Meneely. Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes. Print.

iv) Sequence Data Mining - Guozhu Dong, Jian Pei

v) Data Mining in Bioinformatics - Jason Wang, et. al.

vi) http://www.hindawi.com/journals/tswj/2014/173869/

vii) http://www.sciencedirect.com/science/article/pii/S0888754312000626

viii) https://www.biostars.org/p/43287/