.

PROJECT ASSIGNMENT

DATA MINING

FEATURE-BASED CLASSIFIERS FOR SOMATIC MUTATION PREDICTION

Instructor

Dr.Predrag RADIVOJAC

Authors

Arun RAM(arunsank@umail.iu.edu)

Amol BHAGWAT (arbhagwa@umail.iu.edu)

Mithilesh NANJ (minsrini@indiana.edu)

April 30, 2016

Contents

.

1 Objective and significance

1.1 Objective

The objective of this project is to predict somatic mutation, given a protein sequence using a feature based classifier. This is a neutral vs somatic mutation binary classification problem. Somatic mutations can be divided into passenger and driver mutations. Driver mutations are considered as cancer-causing mutations and an analysis of features affecting somatic vs neutral classification can be a basis of the task of distinguishing between driver and passenger mutations to predict cancer with the mutation information.Citation of Einstein paper [**?**].

1.2 Significance and motivation

In spite of the advancement in health-care technology, cancer deaths have been increasing every year due to increasing variability in the causes of cancer and the randomness associated with predicting the same. Hence, there arises a need of finding a methodology to predict cancer. Driver somatic mutations have been discovered to cause certain variants of cancer and with an exhaustive set of protein sequences/dna, and relevant features, a classification model to predict somatic mutations can be built. Though the problem sounds simple, complexities are involved in both collecting exhaustive protein sequences and finding the most relevant features. The challenges in this field and a dire need for such a model drove us towards this project.

2 Background

2.1 The biology of mutations

A mutation is a permanent alteration of nucleotide sequence of the genome of an organism, virus, or extra-chromosomal DNA or other genetic elements. Mutations result from damage to DNA which is not repaired, errors in the process of replication, or from the insertion or deletion of segments of DNA by mobile genetic elements. Mutations may or may not produce discernible changes in the observable characteristics (phenotype) of an organism. Mutations play a part in both normal and abnormal biological processes including: evolution, cancer, and the development of the immune system, including junctional diversity.

Mutation can result in many different types of change in sequences. Mutations in genes can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely. Mutations can also occur in non-genic regions. One study on genetic variations between different species of Drosophila suggests that, if a mutation changes a protein produced by a gene, the result is likely to be harmful, with an estimated 70 percent of amino acid polymorphisms

that have damaging effects, and the remainder being either neutral or marginally beneficial. Due to the damaging effects that mutations can have on genes, organisms have mechanisms such as DNA repair to prevent or correct mutations by reverting the mutated sequence back to its original state.

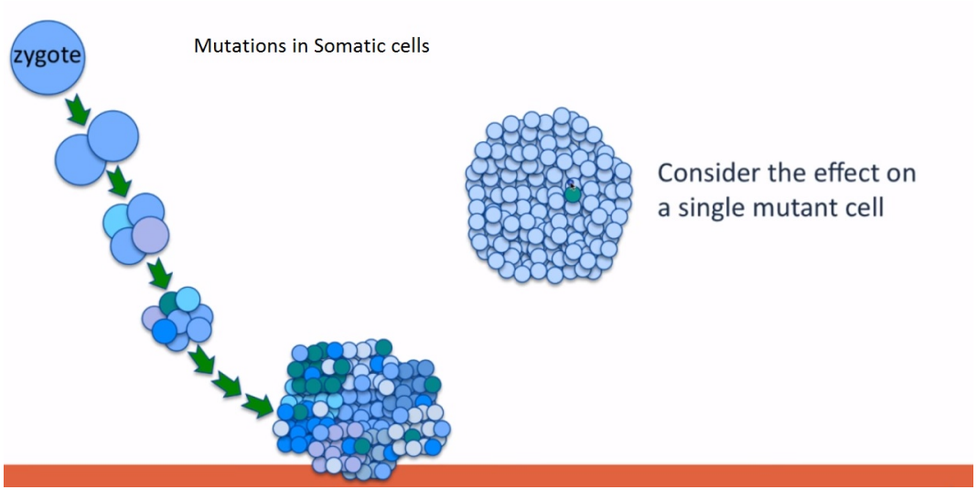


Figure 1: Depiction of mutations in Somatic cells

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. Visualization can be seen in 1

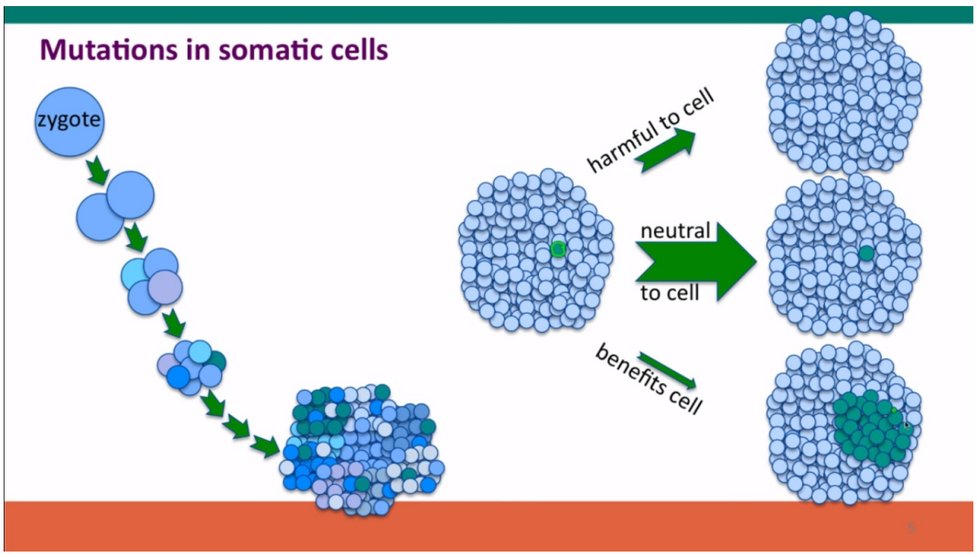


Figure 2: In chronological order, cells that are harmless, neutral and harmful to the organism

As we 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. This can be seen from figure 5

Proteins are the building blocks of life. A protein sequence is a combination of amino acids and the variation in composition of amino acids define the cellular functionality in our body. Driver somatic mutations causes irreversible changes in composition of amino acids and thus the functionality and growth of tumour cells leading to cancer.

2.2 Related work

Works related to somatic mutations include discovering patterns in somatic mutations[**?**], sequence analysis to track patterns of somatic mutations[**?**], somatic mutation behaviour in human genome[**?**]. They have dealt with the behaviour of somatic mutations and their relation to cancer causing genomes.

2.3 Our work

While the literature concentrates on sequential analysis, pattern discovery or a state space model on genome data to detect cancer causing mutations or examine the behaviour of somatic mutations, our work focuses on predicting somatic mutations by a feature based model. Relevant features are considered predictors (which will be explained in later sections) and class is predicted which is either neutral or somatic mutation.

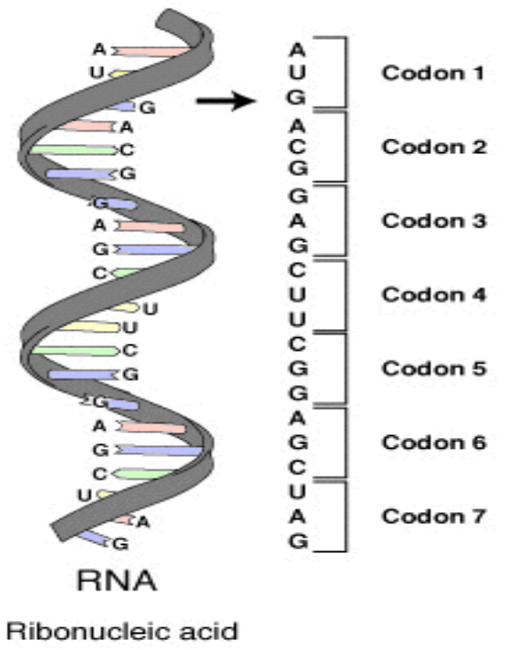
3 Data

3.1 Data description

A nucleic acid sequence is a succession of letters that indicate the order of nucleotides within a DNA (using GACT) or RNA (GACU) molecule. By convention, sequences are usually presented from the 5' end to the 3' end. For DNA, the sense strand is used. Because nucleic acids are normally linear (unbranched) polymers, specifying the sequence is equivalent to defining the covalent structure of the entire molecule. For this reason, the nucleic acid sequence is also termed the primary structure.The sequence has capacity to represent information. Biological deoxyribonucleic acid represents the information which directs the functions of a living thing. Representation and sample data of a sequence are shown in figure 21

3.2 Data Source

Peptide sequence, or amino acid sequence, is the order in which amino acid residues, connected by peptide bonds, lie in the chain in peptides and proteins. The sequence is generally reported from the N-terminal end containing free amino group to the C-terminal end containing free carboxyl group.



(a) RNA Structure

(b) RNA Sequence

Figure 3: RNA (DNA will be similar with a double strand structure and T instead of U in sequence

Peptide sequence is often called protein sequence if it represents the primary structure of a protein. There are a total of 20 amino acids, their categories and representation are given in table 1

4 Model Design

Protein sequence data along with the neutral mutation information from dbSNP was provided by Jose. This corresponds to the data for one of the classes to be predicted - Neutral mutations. Neutral mutations are not harmful and hence are not of interest to the scope of this project. Data for the other class - somatic mutations were obtained from COSMIC (Catalogue Of Somatic Mutations In Cancer) data set, publicly available (http://cancer.sanger.ac.uk/cosmic).

Description	Representation
Any amino acid, or unknown	All
Aspartate derivatives	D, N
Glutamate derivatives	E, Q
Hydrophobic	V, I, L, F, W, Y, M
Aromatic	F, W, Y
Aliphatic	V, I, L, M
Small	P, G, A, S
Hydrophilic	S, T, H, N, Q, E, D, K, R
Positively charged	K, R
Negatively charged	D, E

Table 1: Dimensionality Reduction Results

4.1 Data pre-processing

Since data were from different sources, the following steps were followed to feed them into the models

- Protein sequences were the focus of the project and not genome, but the COSMIC data with somatic mutations had DNA sequences
- DNA sequences were translated into protein sequences using the codon transformation look up. For example the sequence, CGATTTCCT ——> RFP. The translation look up used is given in figure 4
- Mutation positions in the respective protein sequences were also derived
- Mutation in DNA sequence doesn't have to be a mutation in protein sequence since nucleotide to amino acid can be a many to one mapping. We filtered out such occurrences since we are interested only in amino acid mutations
- A string exact matching was done between neutral and somatic mutation data sets to check for common sequences. The common sequences could be grouped under the same ID and it can be made sure that these ids don't go into the same group. Since the COSMIC data didn't have any Ids and also since the string match was only 2%, this was then ignored
- Number of data objects in somatic mutated sequence (COSMIC) (2.7 million records) was 15 times greater than that of neutral mutations (179K records). To avoid a class imbalance problem, under-sampling on COSMIC data was done i.e, records from COSMIC was randomly sampled to match that of neutral data in order to have a balanced class distribution

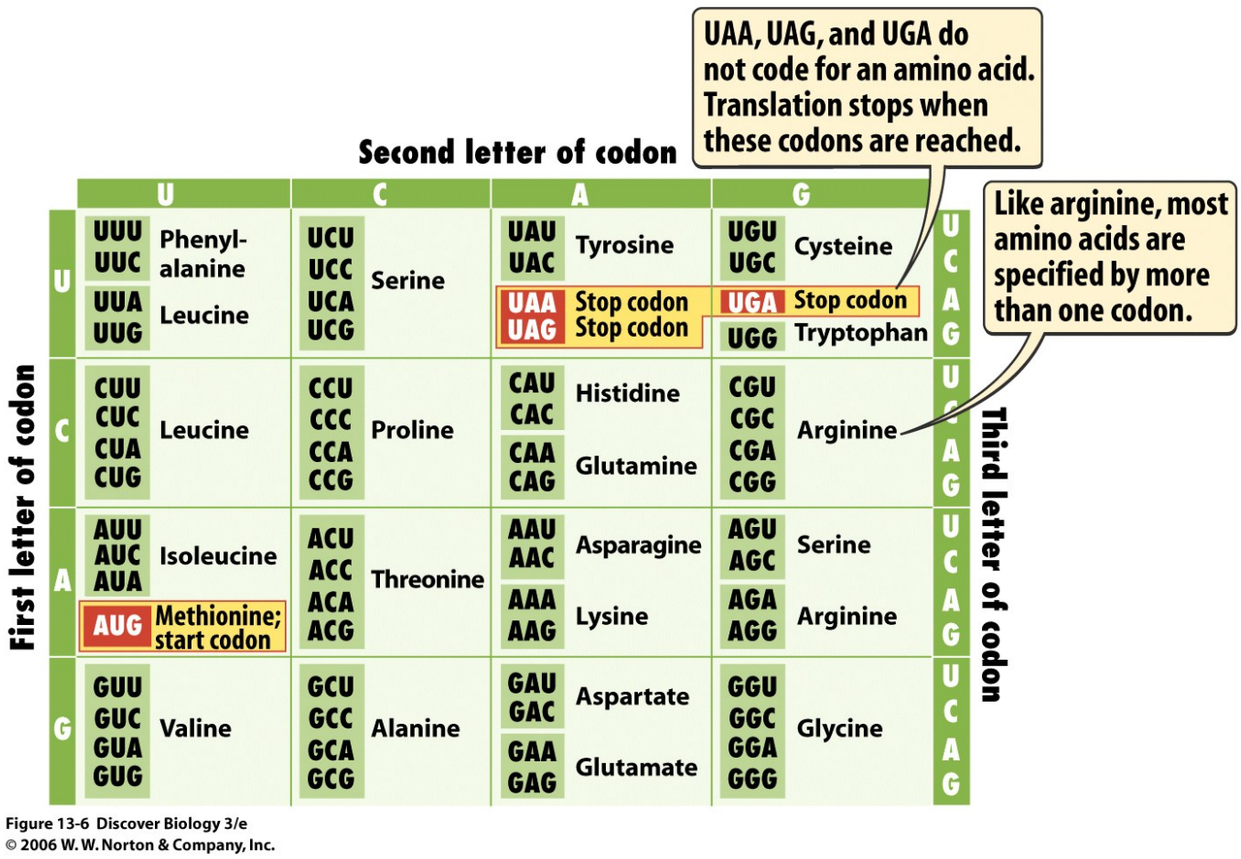


Figure 4: DNA to Amino Acid conversion codons

4.2 Feature engineering

Features which could best represent the mutation were derived from the sequence and mutation information. Following are the features that were extracted and their descriptions

1. Mutation windows - A window around the mutated amino acid was taken and the count of amino acids present in the window apart from the mutated amino acid were the features. Three window sizes were taken - 5,7 and 9. An example for a window size 5 is shown in 2, where C is the position of mutation

2. For a particular window size there will be 20 features which are the counts of amino acids in the window. For the above example, 4 of the 20 amino acids have a value of 1 (count) and the rest of the features (amino acids) will have a value of zero. The same logic goes for window sizes 7 and 9 which makes 60 features

3. One can sense that the data set will be sparse since for a single data object, there can only be a maximum of 4 + 6 + 8 = 18 out of 60 columns can be populated and others will be zero.

4. Entropy for each window size is calculated. This gives the average information that each amino acid in the window offers. These entropies are considered as features because this can be used to distinguish if it is a low complexity region (ex. CCCCC). The equation is given as

$$-\sum_{i=1}^{n} P(X_i) log_2(P(X_i))$$

where P(X) is the fraction of occurrence of a particular amino acid in the window. This accounts another 3 features for the respective window sizes

P	V	C	G	R

Table 2: Window size 5

5. Nature of the amino acids in the window also accounts for the functional changes. This is captured in two features - aromatic amino acid count and charged amino acid count. Each for 3 window sizes, these account for another 6 features

6. Average Hydrophobicity of amino acids in the window was was calculated. Hydrophobicity is defined as the ability of amino acid to restrict the mobility of water molecules thus decreasing the translational and rotational entropy of water. This is included for each window size, adding up 3 features

7. BLOSUM matrix is a mutation matrix which contains likelihood values of a particular amino acid mutating into another. This is a way of encoding our mutated amino acid and allowing the model to learn it. In total there are 73 predictor variables in the data set

4.3 Data definitions

The features in the data set are derived as explained in the previous section in order to predict class 0/1. 0 - Neutral mutations, 1 - Somatic mutations.

4.4 Methodology

The methodology followed is given in the process chart figure 5.

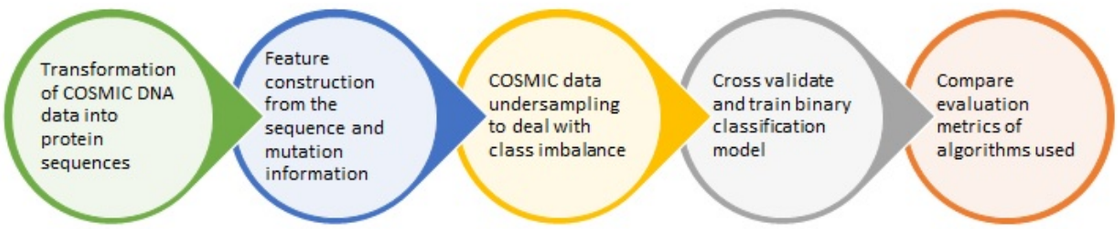


Figure 5: Methodology

The algorithms used in our methodology for classification are Support Vector Machine (SVM), Random forest and logistic regression. SVM has a great advantage when dealing with sparse data

set and we expect this to beat other algorithms, one ensemble method - Random forest and a basic binary classifier Logistic regression.

4.5 Model evaluation

The training set is then subject to 10-fold cross validation. Since it is a binary classifier, accuracy, precision, recall, Area under ROC curve were decided to be used for model evaluation. Since this is a medical science problem, analysis of confusion matrix can be used to evaluate the model. Also, somatic mutations are of interest to this project and hence a precision recall curve is also evaluated since it answers the question of what is the probability that the class is 1 given that our classifier classifies it as 1. In this case, classifying somatic as a neutral mutation has the greatest impact and this can be used to compare the algorithms along with the ROC curve.

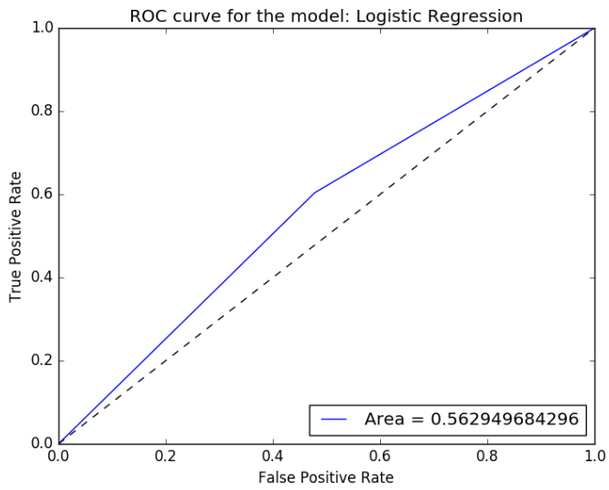
5 Results

The nature of this problem makes it complex in the sense of Three algorithms were tried on this data with a 10-fold cross validation. We wanted to try a diverse set of algorithms and compare the results and hence we chose SVM which is a non-probabilistic kernel type classifier, random forest which is an ensemble decision tree method and logistic regression which is a linear functional classifier. The evaluation metrics and the tuning parameters for each algorithm are explained below.

5.1 Logistic Regression

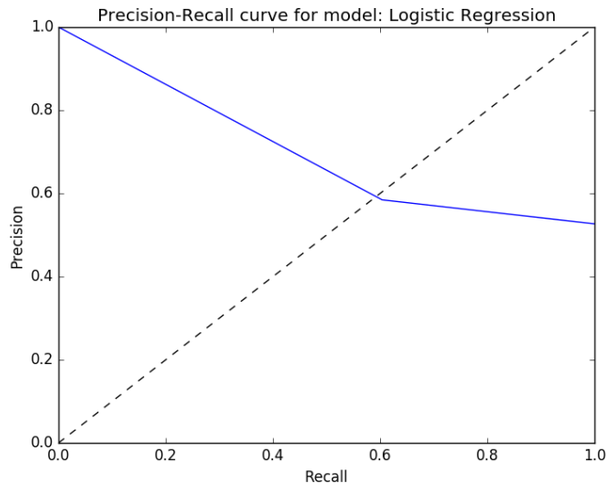
Since it is a binary classification problem, the two simplest probabilistic models that can be used to predict are naive Bayes and Logistic regression. Logistic regression was tried. It assumes a linear relationship between the logit of the variables and the label. Results from this model were taken as the baseline for comparison as more advanced algorithms were used. The ROC curve and precision recall curves for this model are given in figures 6a and 6b

RegressionROC.png



(a) ROC curve

RegressionPR.png



(b) Precision Recall Curve

Figure 6: ROC, Precision Recall Curve for Logistic Regression

		Predicted Class	
		Neutral	**Somatic**
Actual Class	**Neutral**	93,759	85,808
	Somatic	79,248	1,20,752

Figure 7: Confusion matrix - Logistic regression

Figure 8: Confusion bubble - Logistic regression

Observations:

- Area Under Curve - ROC is 0.56. It can be seen that it is better than a random classifier which will have an area of 0.5
- Precision recall in this model for both the classes are almost the same (results are compared in the next section)
- Precision of 0.56 suggests that out of the total somatic mutations predicted, 56% are actually somatically mutated
- This performance can be expected from an algorithm as simple as logistic regression. It is only slightly better than a random classifier
- Confusion matrix is given in figure 7
- It can be seen that somatic mutations are more accurately predicted, from the greater number of true positives (somatic mutations are positives and neutral mutations are negatives)
- Also, with a higher false positives, it suggests that the model classifies neutral mutations as somatic mutations with a higer probability. Precision and recall are presented in the comparison section

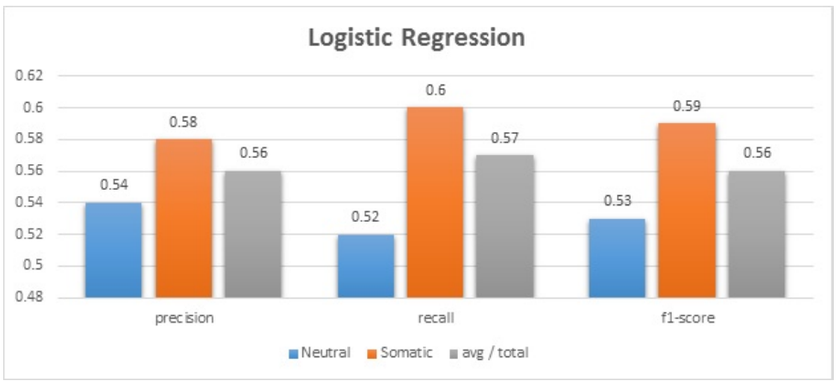


Figure 9: Evaluation metrics - Logistic regression

	precision	recall	f1-score
Neutral	0.54	0.52	0.53
Somatic	0.58	0.6	0.59
avg / total	0.56	0.57	0.56

Figure 10: Evaluation metrics - Logistic regression

5.2 Random Forest

In order to implement an ensemble method for classification, random forest was used. Random forest builds n trees and bootstraps the data and for classification, mode of n classes is assigned as the class for that record. 100 trees and gini impurity for splitting were used as parameters. The ROC curve and precision recall curve are shown in figures 11a and 11b

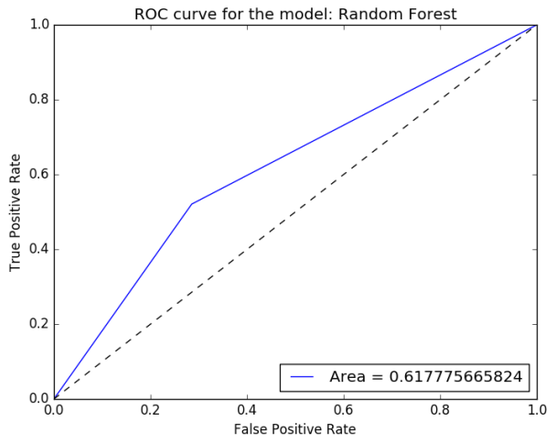
Observations:

- Area Under Curve - ROC is 0.62. It can be seen that it is better than the logistic regression (0.56) and random classifier which will have an area of 0.5
- Precision recall in this model for both the classes vary (results are compared in the next section).
- Random forest model has a better Precision, Recall and f1 score than that of logistic regression. Precision of 0.62 suggests that out of the total somatic mutations predicted, 62% are actually somatically mutated
- Confusion matrix is given in figure 12
- It can be seen that neutral mutations are more accurately predicted, from the greater number of true negatives
- Another difference is that the false positives is 40% less that than of logistic regression

5.3 Support Vector Machines

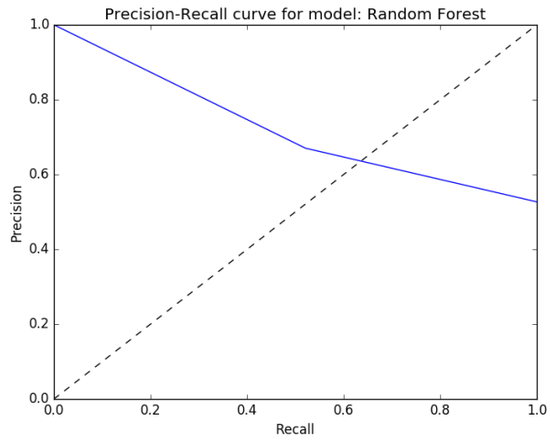
SVM's advantage on a sparse data set is well known. This was exploited in our project since the data set we had was of sparse nature. Both linear and polynomial kernel of degree 3 were used.

Forest ROC.png



(a) ROC curve

Forest PR.png



(b) Precision Recall Curve

Figure 11: ROC, Precision Recall Curves for Random Forest

Linear Kernel

The ROC curve and precision recall curves for SVM with linear kernel are shown in figures 21a and 21b

Observations:

- Area Under Curve - ROC is 0.564. It can be seen that it is the same as logistic regression but better than a random classifier which will have an area of 0.5
- Precision recall in this model for both the classes do not vary (results are compared in the next section)
- Confusion matrix is given in figure 12 which
- It can be seen that somatic mutations are more accurately predicted, can be seen from the greater number of true positives
- Also, with a higher false positives, it suggests that the model classifies neutral mutations as somatic mutations with a higher probability

		Predicted Class	
		Neutral	**Somatic**
Actual Class	**Neutral**	1,28,401	51,166
	Somatic	95,678	1,04,322

Figure 12: Confusion matrix - Random Forest

Figure 13: Confusion bubble - Random Forest

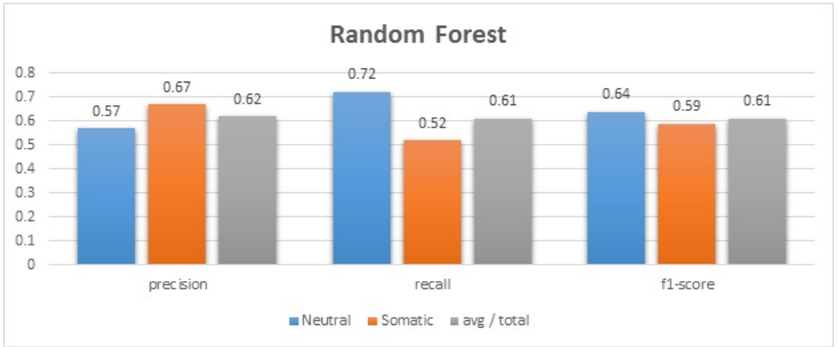
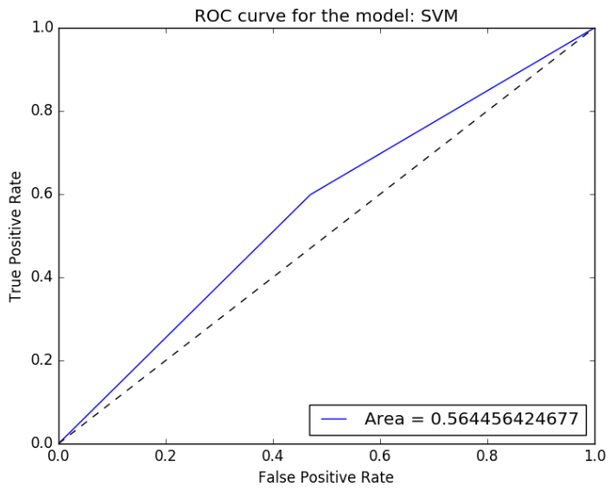


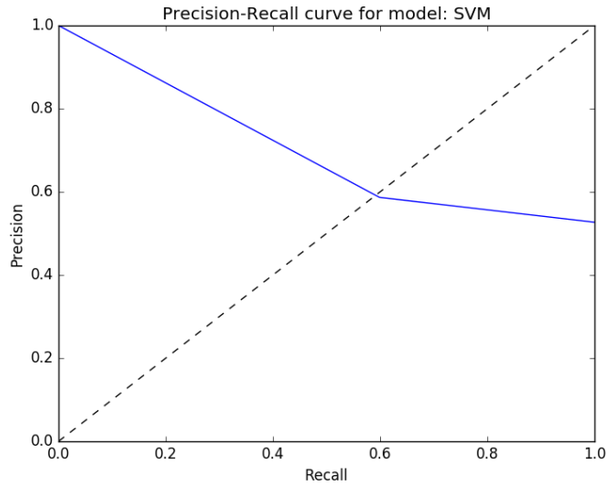
Figure 14: Evaluation metrics - Random Forest

	precision	recall	f1-score
Neutral	0.57	0.72	0.64
Somatic	0.67	0.52	0.59
avg / total	0.62	0.61	0.61

Figure 15: Evaluation metrics - Random Forest



(a) ROC curve



(b) Precision Recall Curve

Figure 16: ROC, Precision Recall Curve for linear SVM

Polynomial Kernel

The ROC curve and precision recall curves for SVM with polynomial kernel are shown in figures 21a and 21b. A polynomial kernel with degree 3 was used.

Observations:

- Area Under Curve - ROC is 0.564. It can be seen that it is the same as logistic regression and linear SVM but better than a random classifier which will have an area of 0.5
- Precision recall in this model for both the classes do not vary (results are compared in the next section)
- Confusion matrix is given in figure 12 which
- It can be seen that neutral mutations are more accurately predicted, can be seen from the greater number of true negatives
- Also, with higher false negatives, it is clear that the model classifies somatic mutations as neutral mutations with a higher probability

		Predicted Class	
		Neutral	**Somatic**
Actual Class	**Neutral**	95,180	84,387
	Somatic	80,228	1,19,772

Figure 17: Confusion matrix - SVM

Figure 18: Confusion bubble - SVM

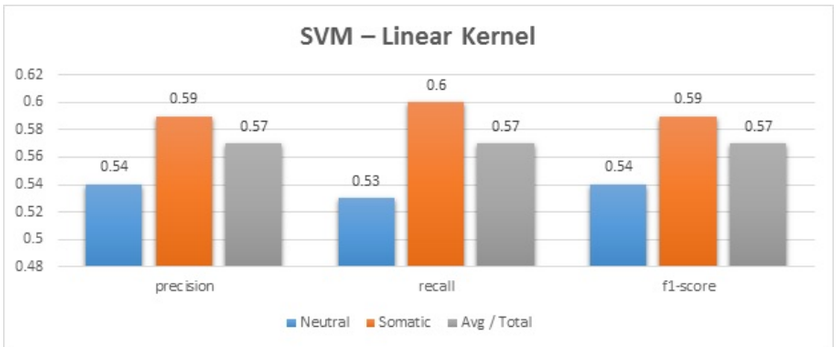
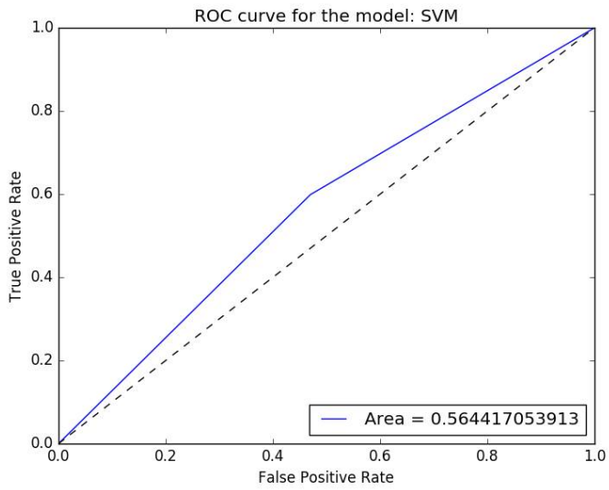


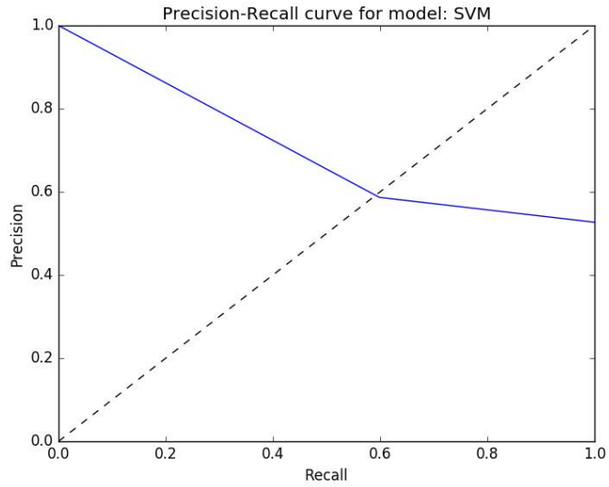
Figure 19: Evaluation metrics - Linear SVM

	precision	recall	f1-score
Neutral	0.54	0.53	0.54
Somatic	0.59	0.6	0.59
Avg / Total	0.57	0.57	0.57

Figure 20: Evaluation metrics - Linear SVM



(a) ROC curve



(b) Precision Recall Curve

Figure 21: ROC, Precision Recall Curve for polynomial SVM

Blosum matrix from our own data was constructed to see the probabilities of mutations that were present in our data and the following are the observations from that

- The mutation E to K has the maximum probability of occurrence with a probability of occurrence of 0.56
- There were numerous mutations with the lowest probability of 0.01

		Predicted Class	
		Neutral	**Somatic**
Actual Class	**Neutral**	1,50,309	51,621
	Somatic	77,432	1,00,205

Figure 22: Confusion matrix - SVM polynomial kernel

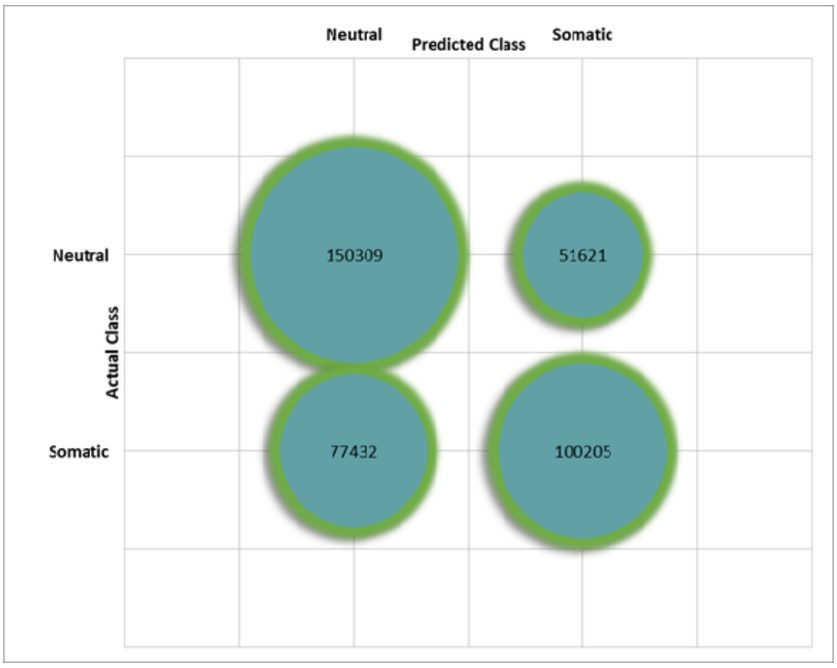
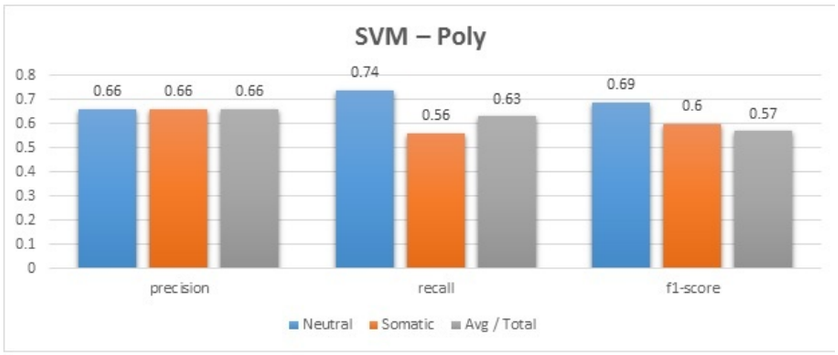


Figure 23: Confusion bubble - SVM Polynomial kernel



Poly.jpg

Figure 24: Evaluation metrics - polynomial SVM

	precision	recall	f1-score
Neutral	0.66	0.74	0.69
Somatic	0.66	0.56	0.6
Avg / Total	0.66	0.63	0.57

Figure 25: Evaluation metrics - polynomial SVM

6 Conclusion

- Distinguishing between neutral and somatic mutations is an ongoing complex research problem due to the randomness of mutation patterns. This is because somatic mutations occur as randomly as neutral mutations
- Our work is a basic step towards accurate classification of somatic and neutral mutations
- With this iteration, we were able to learn about the features and this model gave an idea how different algorithms work on this kind of data and what other features can be used to classify
- Since there is a vast set of derived features available, a different set of features can be used to boost the AUC ROC of the model
- It is an ongoing problem, there is scope for improvement in our model
- Certain measures can be taken to improve the evaluation metrics. Instead of ignoring IDs by exact match, sequences can be aligned and a similarity index can be calculated and sequences with a high similarity can be given the same ID
- This can then be used to cross validate
- PCA can be used for feature selection and the model can have feature selected predictors
- Other features and methods can be explored and implemented to improve accuracy. Also, other data sources can be tried
- In future, with the knowledge gained from this project and a deeper domain knowledge, a more accurate model can be developed
- Among the somatic mutations, cancer causing mutations are called driver mutations and there are other mutations which do not cause cancer which are called passenger mutations
- This can be a base to build a model that can distinguish between driver and passenger mutations

7 Individual Work

- Team member :Arun Ram Sankaranarayanan
- Initial study of the Cosmic dataset
- Initial study of neutral dataset
- Preprocessing of the cosmic dataset
- Calculation of average hydrophobicity
- Calculation using blosum matrix

- Team member :Arun Ram Sankaranarayanan
- Calculation of probability matrix from the cosmic and neutral dataset
- Calculation of count matrix from the cosmic and neutral dataset
- Creation of random forest model
- Visualization of the models and documentation
- Team Member - Mithilesh
- Project idea
- Project background study
- Preprocessing of the cosmic and neutral dataset
- Calculation of aromatic and charged
- Improvisation and inclusion of ideas
- Creation of SVM model
- Creation of logistic model
- Documentation of the predominant part of the report
- Team member : Amol
- Programming and analysis for:
- Recognizing possible flaws and doing preprocessing on the datasets.
- Pairing mutation data with sequences and ensuring correctness of mutation data against the sequences.
- Using the data for and writing code for preliminary feature extraction, including protein counts, residues, and entropy for differing window sizes

8 References

[1] Pan, Y., Karagiannis, K., Zhang, H., Dingerdissen, H., Shamsaddini, A., Wan, Q., . . . Mazumder, R. (2014). Human germline and pan-cancer variomes and their distinct functional profiles. Nucleic Acids Research, 42(18), 11570-11588. doi:10.1093/nar/gku772

[2] GENES AND CHROMOSOMES: Mutation. (1988). Genome, 30(7), 138-168. doi:10.1139/g88-165

[3] GENES AND CHROMOSOMES: Mutation. (1988). Genome, 30(7), 138-168. doi:10.1139/g88-165

[4] http://bioinformatics.oxfordjournals.org/content/29/12/1504.full.pdf+html

9 Link to data and code

https://iu.box.com/s/pgvcck84g8j7wpe90ugzzouebh68hj4h