Scientific Research Publishing

J. Biomedical Science and Engineering, 2017, 10, 28-44
http://www.scirp.org/journal/jbise
ISSN Online: 1937-688X
ISSN Print: 1937-6871

Cancer Specific Non-Synonymous Single Nucleotide Polymorphism Prediction in the Context of Haplotype and Protein Interacting Sites

Pakeeza Akram, Li Liao

Computer and Information Sciences, University of Delaware, Newark, Delaware, USA
Email: liliao@udel.edu

How to cite this paper: Akram, P. and Liao, L. (2017) Cancer Specific Non-Synonymous Single Nucleotide Polymorphism Prediction in the Context of Haplotype and Protein Interacting Sites. *J. Biomedical Science and Engineering*, **10**, 28-44.
https://doi.org/10.4236/jbise.2017.105B004

Received: January 8, 2017
Accepted: May 3, 2017
Published: May 10, 2017

Abstract

In this work, we study predicting the effect of non-synonymous SNPs on several cancers. We trained classifiers on both sequential and structural features extracted from the affected genes and assessed the predictions made by the trained classifiers using cross validation. Specifically, we investigated how the prediction performance can be improved by connecting SNPs in the context of haplotype and interacting sites of proteins encoded by affected genes. We found that accuracy was consistently enhanced by combining sequential and structural features, with increase ranging from a few percentage points up to more than 20 percentage points. The results for putting SNPs in the context of interacting sites were less consistent. Compared to individual SNPs, these that appear together in haplotype showed stronger correlation with one another and with the phenotype, and therefore led to significant improvement inprediction performance, with ROC score increased from 0.81 to 0.95. Although some similar effect has been expected for connecting SNPs to interacting sites in proteins, the performance actually got worse. This decrease in prediction accuracy may be caused by the small data set being used in the study, as many affected proteins in the study do not have known interacting sites.

Keywords

Single Nucleotide Polymorphism, Haplotype, Interaction Sites, Prediction, Cancer

1. Introduction

It has been widely accepted that genetic variations can be associated with diseases. Missense non-synonymous single nucleotide polymorphism (nsSNP) is considered as one of the most common type of variation [1]. Missense nsSNP is a variation in which an amino acid in the protein sequence is changed due to a single point mutation. Because of the association between genetic variations and diseases,

there has been active research to identify SNPs and to determine their phenotypic effects, with some reported success in finding the variants as causes to diagnose, treat and prevent complex diseases [1].

Understanding how these nsSNPs affect protein function remains a critical task. Protein-Protein interaction sites have been considered as a hotspot for nsSNP associated with diseases [2]. In order to unveil genetic variations and functional effect on a protein, multiple methods have been developed, such as enzyme activity prediction [3] [4], detection of disease potential of a SNP [5]. And recently, the computational alanine scanning method is developed to study SNPs effect on protein-protein interaction, essentially by replacing every single residue with alanine tosee the effect on protein by estimating free energy change between the wild and the mutated one [6] [7] [8] [9] [10]. Another recent work has been done for disease associated nsSNPs on protein-protein interactions by investigating the change in binding energy using force field and electrostatic calculation [11].

While most methods have primarily focused on either using sequence based properties such as conservation score alone like SIFT [12] or using only structure based properties such as PoPMuSiC [13], recently there are attempts at hybrid approaches for SNP prediction, such as Polyphen 2, which have showed promising prediction results as compared to using sole properties of structure or sequence [14]. It has also been reported that individual SNPs and haplotypes have different effect on the protein function [15]. In certain cases, it has been found that, with the presence of two SNPs, the disease-causing SNP becomes recessive and does not exert effect on protein function [15]. Despite of the pro- gress, accurate prediction of effect of nsSNP on PPI leading to specific diseases remains a major challenge.

In this paper, we study predicting the effect of non-synonymous SNPs on several cancers, acute myeloid leukemia, breast cancer, colorectal cancer, and esophageal cancer, particularly in the context of haplotype and interaction sites. We formalize the prediction of SNP effects on diseases as a classification problem and then apply machine learning techniques, including support vector machines (SVM) and random forest (RF), to learn from training examples and to classify unseen SNPs. Our comprehensive comparative analysis of different classifiers using a set of evaluation metrics explores not only the utility of various machine learning methods for this problem but also whether and how prediction of SNP's effect is affected for genetic variations by their presence at interacting sites and non-interacting sites of the protein, or for individual SNPs versus SNPs as haplotype associated with a specific disease.

2. Methods

As mentioned above, we formalize the prediction of SNP's effects on proteins associated with specific diseases as a classification problem and adopt supervised learning strategy. Specifically, two powerful classifiers, random forest [22] and support vector machines [23], are selected for this study. For SVM, 3 different kernels were adopted and assessed: Linear, Radial Basis Function

$K^G(x,x') = \exp(-\|x-x'\|^2/c)$ where the values for C = 3.46 and Polynomial $K^P(x,x') = (\langle x,x'\rangle + 1)^d$ with degree d = 2 was applied. These values of C and degree of polynomial d were optimized by using Opunity 1.1.1, a python package.

Features, both sequential and structural, of proteins encoded by genes with SNPs that are believed to be relevant for the phenotypic properties are collected and quantified for use as input vector **x** to the classifier. Specifically, for this study, we are interested in two types of phenotypic properties: detrimental or polymorphic, corresponding to the output y of the binary classifier, namely, y = 1 for detrimental and 0 for polymorphic. The classifier is to learn the actual mapping from input to output: y = F(**x**), with a hypothesis function H(**x**, ɵ), where ɵ collectively represents the parameters of the classifier, for example the degree d of a polynomial kernel for SVM. The classifier is trained to minimize the empirical error

$$\min \{\Sigma_{i=1 \text{ to } n} | F(\mathbf{x_i}) - H(\mathbf{x_i}, \theta) |\} \quad (1)$$

for a set of n training examples $\mathbf{x}_i$, i = 1 to n, whose phenotypic property $y_i = F(\mathbf{x_i})$ is known. Once the classifier is trained, it is used to make prediction / classification on unseen data, *i.e.*, SNPs whose phenotypic property is not known a prior.

Feature selection plays a critical role in ensuring effective learning and reliable prediction. It has been known that mutations that occur at the interface between interacting proteins are more likely to cause detrimental effect as compared to present on other sites. Also, previous studies suggest that haplotype may have influence on whether a particular SNP may or may not manifest its phenotypic effect. Therefore, in this study, we are particularly interested in predicting the effect of non-synonymous SNPs on four types of common cancers in the context of SNPs being on protein interaction sites or within a haplotype.

The pipeline developed for this study consists of steps for data collection, feature characterization/quantification, classifier training, testing and evaluation, as shown in **Figure 1**. Detail for each step is given in the following subsections.

2.1. Data and Feature Characterization

SNPs and phenotypic effect for the four different types of cancers-acute myeloid leukemia (MIM # 601626), breast cancer (MIM#114480), colorectal cancer (MIM#114500) and esophageal cancer (MIM#114480) are collected from OMIM, one of the biggest databases which provides detailed information about phenotype-genotype relation [16].

To determine whether SNPs occur at protein-protein interaction sites, we used STRING database to identify the interaction sites for the affected proteins (*i.e.*, the gene products) [17]. For Acute Myeloid Leukemia, 16 genes are in-

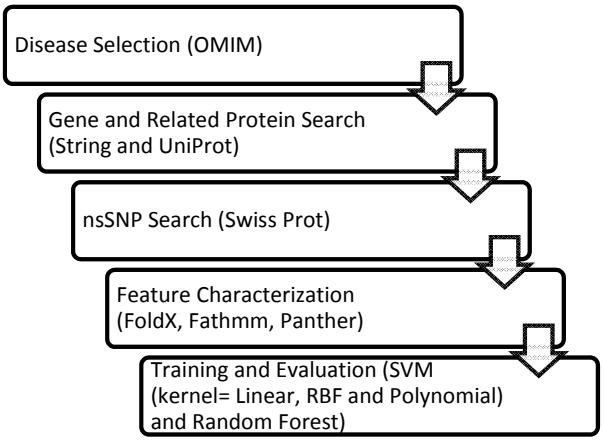

Figure 1. Pipeline constructed for nsSNP prediction starting from gene collection to classification estimation.

volved, which result in 171 proteins that have certain interactions with each other. Due to unsolved 3D protein structures the set is reduced to 111 proteins. There are several databases which provide SNP data, including SwissProt and dbSNP. For this study we used SNP from SwissProt database [18] because of its large collection as compared to other databases. The queries to SwissProt identified 1399 nsSNP for these 111 proteins. The same data collection protocol is used for the other three cancers as well. After filtering with required protein structural as well as sequence properties, the final data set consists of 4056 SNP's in total, as listed in **Table 1**.

Using these nsSNP, feature vector was constructed using several properties of both sequence and their respective structure. FoldX was used to calculate parameters which are important for protein stability [19]. It provides several important features along with the calculation of total energy for the mutant and the wild type protein. Panther software calculates Substitution Position-Specific Evolutionary Conservation (subPSEC) Scores and it is based on hidden Markov model (HMM). It was used to collect subPSEC score. Fathmm was used to calculate HMM cancer-specific pathogenicity weights [21]. In total 21 features were collected and all these features are shown in supplementary data S1.

We also collect haplotype data for genes associated with Acute Myeloid Leukemia. A haplotype is considered as set of polymorphic, which are inherited together. It is referred to a combination of alleles or a set of SNP that are found on the same chromosome [15]. To collect haplotype information two databases were used in this study. One is HapMapProject and the other is UCSC genome browser [25] [26]. HapMap Project has a wide range of SNPs, which are collected from dbSNP. Since our dataset consists of SNPs collected from SwissProt, to collect as many as haplotype data, we incorporate UCSC genome browser, which provides gene based common allele variants taken from 1000 genome project [27].

2.2. Cross-Validation and Evaluation

To assess the prediction performance, we adopt the widely accepted cross-vali-

Table 1. Data distribution for cancer type representing polymorphic and detrimental SNP's.

Cancer	Polymorphic	Detrimental	Total
Acute Myeloid Leukemia	1131	268	1399
Breast Cancer	1087	145	1232
Colorectal Cancer	983	131	1114
Esophageal Cancer	961	94	1055
Total	3473	583	4056

dation scheme. Specifically, we used 10-fold cross-validation. The data is randomly split to 10 equal-sized subsets, and one set is reserved for testing and the remaining 9 subsets are combined into a training set to train the classifier. This process is repeated 10 times, with each subset being used as test set once and the average performance from 10 runs is reported. We used some commonly used measurements to report the performance, which includes accuracy, precision, recall, F1 score, and MCC, defined as follows.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP stands for true positive when a SNP is correctly predicted as detrimental, TN for true negative when a SNP is correctly predicted as polymorphism, FP for false positive when a SNP is incorrectly predicted as detrimental; and FN for false negative when a SNP is incorrectly predicted as polymorphism.

We also evaluate the performance using receiver operating characteristic (ROC) curve and Receiver operating characteristic (ROC) score.ROC is a graphical representation that illustrates the performance of a binary classifier system. The plot is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or recall while false-positive rate is also known (1 − specificity) [28].

3. Results and Discussions

In this study we carried out comprehensive comparative analysis of predicting SNPs effects on four types of cancers. Specifically, we examined the following four different scenarios:

1. Comparison using structural properties only, or sequence properties only or combine effect of both properties using different classifiers;
2. Specific cancer SNP's prediction or collection of cancers SNP's prediction;
3. SNP's prediction for residues at interacting sites or non-interacting sites;
4. SNP's prediction for SNPs within haplotype or individual SNP's.

Note that, due to data collection issues, the last two types of analysis were only performed for Acute Myeloid Leukemia.

3.1. Comparison Using Structural Properties Only, or Sequence Properties Only or Combine Effect of Both Properties Using Different Classifier

For the 4056 SNP's listed in **Table 1**, three different datasets were generated. All three datasets have the same number of instances but different dimensionality of the feature vector. First dataset had 3 (sequential) features in it, second dataset had 18 (structural) features and the last dataset had all 21features in it. Receiver operating characteristic (ROC) score was calculated for 10-Fold cross validation and the mean of those score is represented in **Table 2** and **Figure 2** respectively.

The results clearly show that using structural and sequence based features together for SNP Prediction provides better results as compared to individual protein properties. It also suggests that hybrid features provide better results for any combination of features used. It also shows that random forest performs better among other classifiers used in this task.

3.2. Specific Disease SNP's Prediction or Collection of Diseases SNP's Prediction

For this task, data was collected for four different cancers that are breast cancer,

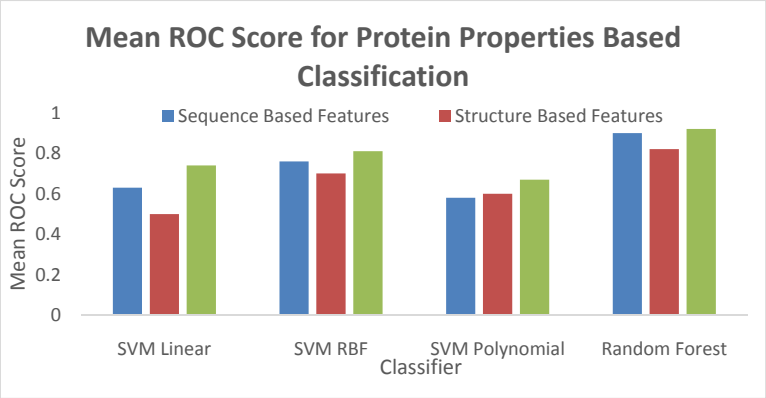

Figure 2. Classifier performance using ROC Score for sequence based, structure based and hybrid protein properties.

Table 2. Mean ROC score for SNP prediction using different classifiers for specific protein based properties.

Classifier	Sequence Based Features	Structure Based Features	Hybrid Features
SVM Linear	0.63	0.5	0.74
SVM RBF	0.76	0.7	0.81
SVM Polynomial	0.58	0.6	0.67
Random Forest	**0.9**	**0.82**	**0.92**

colorectal cancer, esophageal cancer and acute myeloid leukemia, see **Table 1**. It was observed that very few genes, such as TP53, were common for all types of

cancers collected for this study and generally in all types of cancers. It can be seen from **Table 1** that the number of detrimental SNPs is low as compared to the polymorphic SNP's. The difference is almost three times between two types of SNPs. Prediction performance for every classifier for each disease was studied. **Table 3** lists the performance of each classifier on both detrimental as well as polymorphic SNP.

Table 3. Evaluation metric score for each cancer using four different classifiers.

Cancer Type	Classifier	SNP Type	Precision	Recall	F1-Score	Accuracy
Acute Myeloid Leukemia	SVM Linear	Polymorphic	**0.87**	0.91	**0.89**	
		Detrimental	**0.53**	**0.41**	**0.46**	**0.82**
	SVM RBF	Polymorphic	0.84	0.94	**0.89**	
		Detrimental	0.5	0.26	0.34	0.81
	SVM Polynomial	Polymorphic	0.83	**0.96**	**0.89**	
		Detrimental	0.51	0.17	0.26	0.81
	Random Forest	Polymorphic	0.86	0.92	**0.89**	
		Detrimental	0.51	0.35	0.41	0.81
Breast Cancer	SVM Linear	Polymorphic	**0.88**	**0.9**	**0.89**	
		Detrimental	0.13	0.1	0.11	**0.81**
	SVM RBF	Polymorphic	**0.88**	0.91	**0.9**	
		Detrimental	**0.11**	0.08	0.09	**0.81**
	SVM Polynomial	Polymorphic	**0.88**	0.9	0.89	
		Detrimental	0.13	0.11	0.12	**0.81**
	Random Forest	Polymorphic	**0.89**	0.88	0.88	
		Detrimental	0.14	**0.15**	**0.15**	**0.81**
Colorectal Cancer	SVM Linear	Polymorphic	0.88	**0.96**	0.91	
		Detrimental	0	0	0	0.84
	SVM RBF	Polymorphic	0.88	**0.96**	**0.92**	
		Detrimental	0.07	0.02	0.03	**0.85**
	SVM Polynomial	Polymorphic	0.88	0.95	0.91	
		Detrimental	0	0	0	0.84
	Random Forest	Polymorphic	**0.89**	0.94	0.91	
		Detrimental	**0.26**	**0.17**	**0.2**	0.84
Esophageal Cancer	SVM Linear	Polymorphic	0.91	**0.99**	**0.95**	
		Detrimental	0	0	0	0.9
	SVM RBF	Polymorphic	**0.92**	**0.99**	**0.95**	
		Detrimental	**0.44**	**0.07**	**0.13**	**0.91**
	SVM Polynomial	Polymorphic	0.91	**0.99**	0.95	
		Detrimental	0.08	0.01	0.02	0.9
	Random Forest	Polymorphic	0.91	0.98	0.94	
		Detrimental	0.14	0.03	0.05	0.9

The above table represents that SVM RBF performs better for esophageal and colorectal cancer and SVM linear performed better for acute myeloid leukemia,

while all classifiers performed about equally well on breast cancer. It also shows that for polymorphic SNP prediction precision and recall is much better as compared to the detrimental SNPs. This may be attributed to the skewed data distribution. It is also noticeable that in terms of accuracy there is only 1% difference while using different classifiers.

Further, all the cancer types were lumped together to analyze their performance (shown in **Table 4**). It showed that random forest once again performed better. In order to further evaluate predictive power without using a fixed threshold to determine positive versus negative, receiver operating characteristic (ROC) score was calculated for all classifiers using 10-fold cross validation. The mean ROC score is represented in **Figure 3**. Results from mean ROC score show that except for acute myeloid leukemia for each disease random forest provides better score. And in general, all the ROC Scores are above 0.70.

Initially, it was hypothesized that SNP classification for individual disease will be better than that of combine diseases but results reflect the opposite. In order to further investigate couple of tasks were performed. It was noticed that there were six genes which are common and associated with cancer types selected for this study. These common genes were completely removed from data set and classification was performed. Results showed that mean ROC score for all the cases was less than 0.6 (shown in **Figure 3**). It provides a clue that if there is no common gene among diseases than SNP prediction for individual cancer type

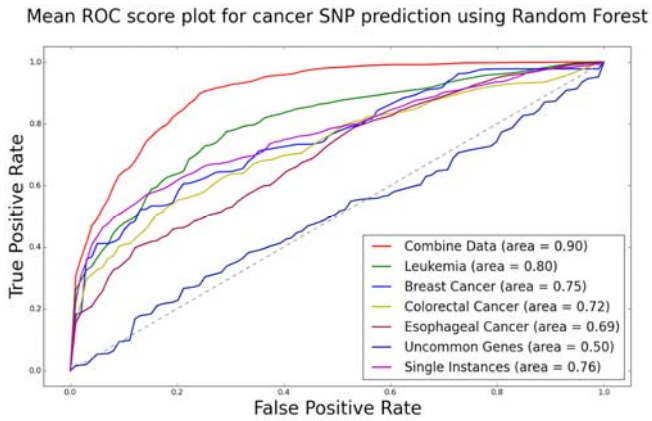

Figure 3. Mean ROC score plot for each cancer type using random forest (best classifier for study).

Table 4. Mean ROC score for balanced and unbalanced collective cancer data

Mean ROC Score	Combine Cancers	Single Instance	Balanced Data
SVM Linear	0.61	0.69	0.88
SVM RBF	0.71	0.73	0.88
SVM Polynomial	0.7	0.72	0.88
Random Forest	**0.9**	**0.76**	**0.9**

will be better but in general almost all the cancers have certain common genes.

Another task was performed to see how training be affected if the combination of all disease SNP without redundancy *i.e.* only single instance of SNP occur in

the final dataset when this gene is shared by more than one cancer type. In this case ROC score was similar to every individual cancer type SNP classification.

It was noticed and mentioned earlier that detrimental SNP are much less in number than the polymorphic SNPs. It produces an unbalanced dataset. To see what impact data would make if the number of detrimental SNP is equal to polymorphic SNPs. Number of SNPs for polymorphic class was reduced and then classification task was performed. It does not show any change in ROC score for best classifier but the F1-score for detrimental SNPs was rapidly increased from 0.45 to 0.86. This change in detrimental SNP evaluation can be seen from **Table 5** as well as from the **Figure 4**. It was noticed that when data is balanced it does not affect polymorphic SNPs but classification of detrimental SNP is significantly improved.

Lastly mean ROC score was calculated using 10-fold cross validation for each classifier and found that random forest provides better results as compared to any other classifier. Note that there is no change in the mean ROC score for best classifier but SVM with its different kernels is performing better.

To assess the statistical significance for the difference between that set of combine cancers and the set of Acute Myeloid Leukemia, a t-test was performed on the ROC score of both datasets using random forest, and p-value is 0.007458. This concludes that random forest performs better than other classifiers when SNP's prediction is done for any type of cancer.

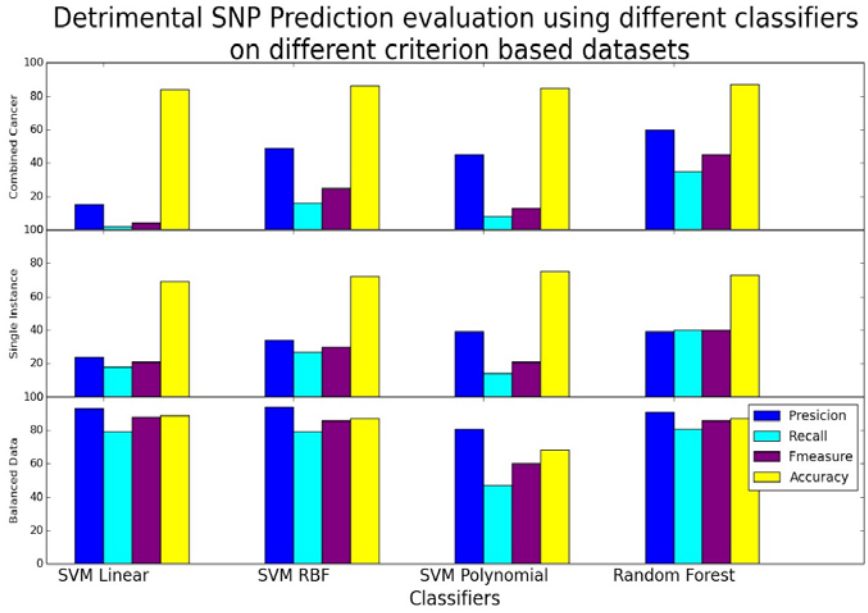

Figure 4. Detrimental SNP evaluation for Combined data, single instance data (non-redundant) and Balanced data. In case of balanced data performance is rapidly improved but in all cases random forest is performing better.

Table 5. Evaluation metric score for combined cancer SNP using four different classifiers.

Cancer Type	Classifier	SNP Type	Precision	Recall	F1-Score	Accuracy
Combine Cancers	SVM Linear	Polymorphic	0.86	**0.98**	0.91	
		Detrimental	0.15	0.02	0.04	0.84

Single Instance	SVM RBF	Polymorphic	0.87	0.97	0.92	
		Detrimental	0.49	0.16	0.25	0.86
	SVM Polynomial	Polymorphic	0.86	**0.98**	0.92	
		Detrimental	0.45	0.08	0.13	0.85
	Random Forest	Polymorphic	**0.9**	0.96	**0.93**	
		Detrimental	**0.6**	**0.35**	**0.45**	**0.87**
	SVM Linear	Polymorphic	0.77	0.84	0.8	0.69
		Detrimental	0.24	0.18	0.21	
	SVM RBF	Polymorphic	0.8	0.85	0.82	0.72
		Detrimental	0.34	0.27	0.3	
	SVM Polynomial	Polymorphic	0.79	**0.93**	**0.85**	**0.75**
		Detrimental	**0.39**	0.14	0.21	
	Random Forest	Polymorphic	**0.82**	0.82	0.82	0.73
		Detrimental	**0.39**	**0.4**	**0.4**	
Balanced Data	SVM Linear	Polymorphic	**0.83**	**0.99**	**0.9**	
		Detrimental	0.93	0.79	0.88	**0.89**
	SVM RBF	Polymorphic	0.82	0.95	0.88	
		Detrimental	**0.94**	0.79	**0.86**	0.87
	SVM Polynomial	Polymorphic	0.63	0.89	0.74	
		Detrimental	0.81	0.47	0.6	0.68
	Random Forest	Polymorphic	**0.83**	0.92	0.87	
		Detrimental	**0.91**	**0.81**	**0.86**	0.87

3.3. SNP Prediction for Residues at Interacting Site or Non-Interacting Site

nsSNP prediction was done at interacting site as well as non-interacting site. 3DID database (release: June 2015) was used to observe presence of a particular residue at interacting site. It was found that among 40 proteins associated with acute myeloid leukemia having solved 3D structure and nsSNP there are only 18 proteins which had information for their interacting and non-interacting residues recorded in the database. Two subsets were created for this problem one having SNPs at interacting residues and the other with SNPs at non-interacting residues. Data distribution is shown in **Table 6**.

Classification prediction was performed using same classifiers. Their performance with reference to precision, recall, F1-Scoreand accuracy is given below in **Table 7**. Data distribution is balanced for both subsets and thus it provides im-

Table 6. SNPs at Interacting Sites versus Non-interacting Sites.

Acute Myeloid Leukemia	Polymorphic	Detrimental	Total
Interacting Site Residue	58	43	**101**

Non-Interacting Site Residue	131	120	**251**

Table 7. Evaluation metric score for SNPs at interacting and non-interacting sites using four different classifiers.

Cancer Type	Classifier	SNP Type	Precision	Recall	F-Measure	Accuracy
Interacting Site Residues	SVM Linear	Polymorphic	**0.68**	0.71	0.7	
		Detrimental	0.6	**0.58**	**0.59**	0.65
	SVM RBF	Polymorphic	0.68	0.66	0.67	
		Detrimental	0.56	**0.58**	0.57	0.62
	SVM Polynomial	Polymorphic	0.58	**0.84**	0.69	
		Detrimental	0.44	0.16	0.24	0.57
	Random Forest	Polymorphic	**0.68**	0.74	**0.71**	
		Detrimental	**0.61**	0.53	0.57	**0.67**
Non-interacting Site residues	SVM Linear	Polymorphic	0.54	**0.56**	0.55	
		Detrimental	0.5	0.49	0.5	0.53
	SVM RBF	Polymorphic	0.54	0.58	0.56	
		Detrimental	0.5	0.47	0.48	0.53
	SVM Polynomial	Polymorphic	**0.59**	**0.89**	**0.71**	
		Detrimental	**0.72**	0.32	0.44	**0.61**
	Random Forest	Polymorphic	0.56	0.56	0.56	
		Detrimental	0.52	0.53	**0.52**	0.55

proved results for both datasets when compared to task one datasets in terms of polymorphic and detrimental prediction.

While the overall performance has been dropped, there is an improved performance for prediction of detrimental SNP's. Further, ROC score was determined for all classifiers for both datasets as shown in **Figure 6**. The upper panel is for all the classifier trained and tested for SNPs at interacting sites and the lower panel is for non-interacting site SNP's. Mean ROC score for SVM RBF and SVM polynomial were same *i.e.* 0.86 for both datasets but in case of non-interacting site residues SVM polynomial is performing better with 0.66 score. It concludes that when overall performance of two datasets is considered SVM polynomial has better performance than any other classifier. Lastly to verify the statistical significance of the performance difference, a t-test was performed on the 10-fold cross validation of SVM polynomial ROC score and it was found that p-value is 0.020197, confirming the statistical significance of the difference.

3.4. SNP Prediction Individual SNPs vs SNPs within Haplotype

In this analysis, we examine predicting SNPs effect in the context of haplotype,

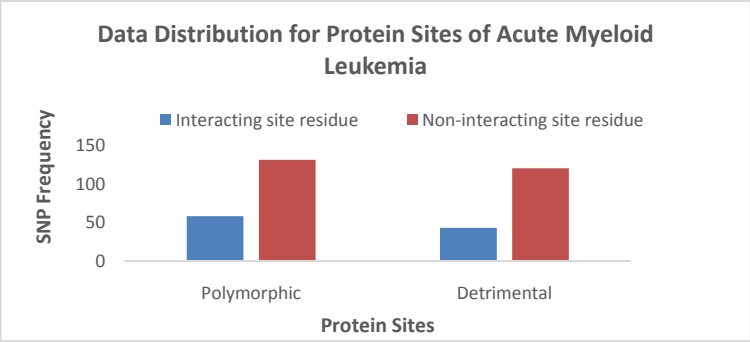

Figure 5. SNP data distribution for acute myeloid leukemia at interacting and noninteracting site of protein

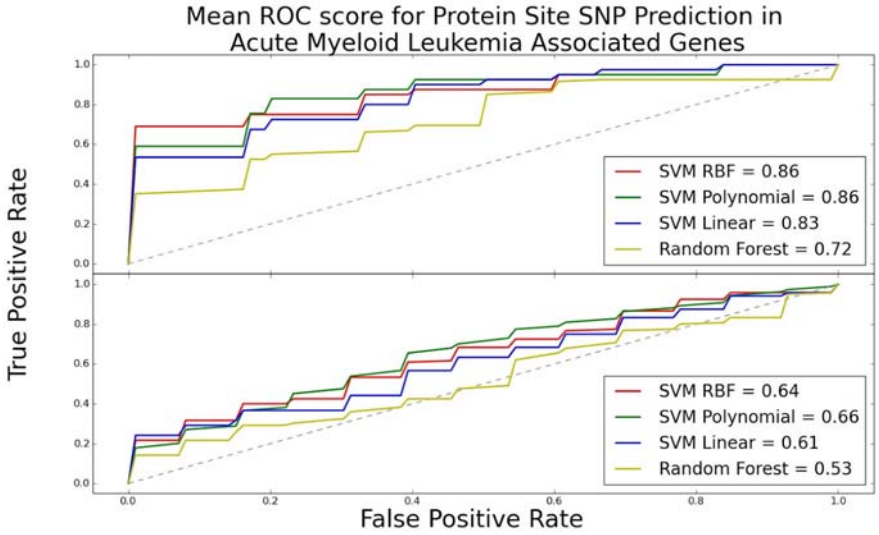

Figure 6. Mean ROC score plot for several classifiers at interacting site (upper plot) and at noninteracting site (lower plot) of protein.

i.e., the prediction of individual SNPs versus SNPs within a known haplotype. The search against database from HapMap Project and the other is UCSC genome browser only identified haplotypes from 14 genes from the gene pool associated with acute myeloid leukemia. Haplotypes were considered in pair only that means each single SNP in haplotype was compared to every haplotype allelic change within same gene including self-replication. In this task, two subsets were generated: one set consists of haplotypes pairs and the other set consists of all individual SNPs associated with genes involved in acute myeloid leukemia. Data distribution for these two subsets is given in **Table 8**.

For training 10-fold cross validation was applied to both datasets using SVM with three kernels and random forest. The results for this classification problem are shown in **Table 9**.

In **Table 9** we can see easily that the best accuracy in predicting haplotype pair is 0.91, a significant increase over 0.82, the best accuracy in predicting individual SNPs. Also, we notice a clear advantage of Random forest for predicting haplotype pairs across the board on all four metrics, whereas SVM Polynomial p e r -

Table 8. Data distribution for haplotype and individual gene in acute myeloid leukemia.

Acute Myeloid Leukemia	Polymorphism	Detrimental	Total
Haplotype Pair	1109	316	**1425**
Individual SNP's	1053	217	**1270**

Table 9. Evaluation metric score for SNPs in haplotype pair or individual SNP using four different classifiers.

Cancer Type	Classifier	SNP Type	Precision	Recall	F1-Score	Accuracy
Haplotype Pair	SVM Linear	Polymorphic	0.85	0.81	0.83	0.74
		Detrimental	0.43	0.5	0.46	
	SVM RBF	Polymorphic	0.88	0.87	0.88	0.81
		Detrimental	0.57	0.6	0.58	
	SVM Polynomial	Polymorphic	0.88	0.88	0.88	0.82
		Detrimental	0.59	0.59	0.59	
	Random Forest	Polymorphic	**0.96**	**0.92**	**0.94**	**0.91**
		Detrimental	**0.75**	**0.88**	**0.81**	
Individual SNP	SVM Linear	Polymorphic	**0.87**	0.91	0.89	0.81
		Detrimental	0.44	**0.35**	**0.39**	
	SVM RBF	Polymorphic	0.85	0.92	0.88	0.8
		Detrimental	0.37	0.24	0.3	
	SVM Polynomial	Polymorphic	0.86	**0.93**	**0.9**	**0.82**
		Detrimental	**0.45**	0.27	0.34	
	Random Forest	Polymorphic	0.85	0.87	0.87	0.79
		Detrimental	0.32	0.3	0.32	

forms slightly better for predicting of individual SNPs. In particularly, it is worth noting that the F1-score for haplotype pair of detrimental phenotype is 0.81 by Random Forest classifier, which is a very impressive performance given that the datasets (**Table 8**) are quite skewed toward polymorphic phenotype and therefore present a greater challenge for correctly predicting the detrimental phenotype. The four metrics used in **Table 9** all depend on a fixed threshold for prediction. ROC curve and score can evaluate a classifier's predictive power and performance without relying on a specific prediction threshold. In **Figure 7**, ROC curves and scores are shown for haplotype SNP pairs (top panel) and individual SNPs (bottom panel). The two key observations from **Table 9** are essential maintained: a) pairing SNPs in haplotype help improve phenotype prediction (ROC score = 0.95, achieved by RF), as compared to predicting phenotype for individual SNPs (ROC score = 0.81, achieved by SVM-RBF); b) while RF generally performs better, SVM-RBF has a slight edge in predicting individual SNPs.

Again, a t-test was performed on ROC scores from the 10-fold cross validation using Random Forest for haplotype pair versus individual SNPs. The p-value is

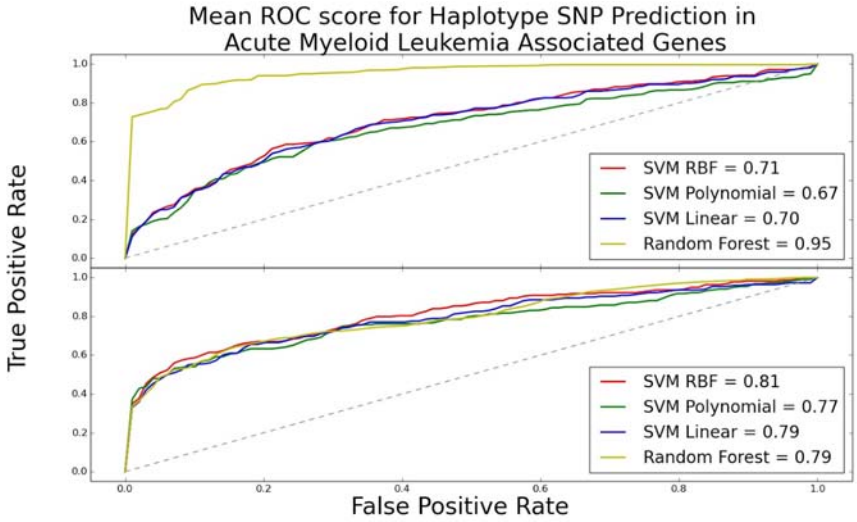

Figure 7. Mean ROC score plot for haplotype pair (upper panel) and individual SNP prediction (lower panel).

7.8×10^{-15}, confirming the statistical significance of the difference.

Overall, it suggests that Random forest is the better classifier for most of the tasks performed in this study. An exception was seen for task 3, where SVM polynomial is providing better results.

4. Conclusions and Future Work

In this work, we carried out comprehensive comparative analysis for predicting SNPs effect associated with four types of cancers, in the context of SNPs being present at protein interacting sites versus non-interacting sites and being paired within a known haplotype versus being unpaired.

Our results confirm that prediction performance is generally improved from using both sequential features and structural features than using them separately. Also, of the two types of classifiers used in the study, random forest outperforms in most cases.

It is found that generic SNP prediction provides better association of particular SNP to be detrimental or polymorphic SNPs as compared to disease-specific SNPs, although this conclusion does not hold if genes associated with one disease are unique from the other disease. While it is expected that prediction performance will be increased by associating SNPs to the interacting sites, the results show instead slight decrease in performance. This decrease in predicting accuracy may be caused by the small data set, as many affected proteins in the study do not have known interacting sites.

Compared to individual SNPs, these that appear together in haplotype showed stronger correlation with one another and with the phenotype, and therefore led to better prediction performance. Haplotype SNP prediction provided most promising results. This could be taken to the next level of improving further accuracy and developing personalized drug. Although currently the haplotype classification and protein site classification was performed for only Acute Myeloid

Leukemia, the same protocol can be adopted to perform similar analysis on other diseases.

Lastly, while this study was performed on cancer diseases only, the same protocol could be applied for the prediction of non-cancerous diseases in order to make this protocol generic for all diseases.

References

[1] Wu, J., Gan, M., and Jiang, R. (2011) Prioritisation of Candidate Single Amino Acid Polymorphisms Using One-Class Learning Machines. *International Journal of Computational Biology and Drug Design*, **4**, 316–331. https://doi.org/10.1504/IJCBDD.2011.044446

[2] David, A., Razali, R., Wass, M.N. and Sternberg, M.J. (2012) Protein-protein Interaction Sites are Hotspots for Disease-Associated Nonsynonymous SNPs. *Hum Mutat* **33**, 359–363. https://doi.org/10.1002/humu.21656

[3] Basit, N. and Wechsler, H. (2011) Prediction of Enzyme Mutant Activity Using Computational Mutagenesis and Incremental Transduction. Advances in bioinformatics. *Adv Bioinformatics,* **2011**, 958129. https://doi.org/10.1155/2011/958129

[4] Lee, T.S. and York, D.M. (2010) Computational Mutagenesis Studies of Hammerhead Ribozyme Catalysis. *J Am ChemSoc*, **132**, 13505–13518. https://doi.org/10.1021/ja105956u

[5] Masso, M. and Vaisman, II. (2010) Knowledge-based Computational Mutagenesis for Predicting the Disease Potential of Human Non-Synonymous Single Nucleotide Polymorphisms. *J TheorBiol* , **266**, 560–568. https://doi.org/10.1016/j.jtbi.2010.07.026

[6] Bradshaw, R.T., Patel, B.H., Tate, E.W., Leatherbarrow, R.J. and Gould, I.R. (2011) Comparing Experimental and Computational Alanine Scanning Techniques for Probing a Prototypical Protein–Protein Interaction. *Protein Engineering Design and Selection*, **24**, 197–207. https://doi.org/10.1093/protein/gzq047

[7] Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., *et al*. (2010) A Method and Server for Predicting Damaging Missense Mutations. *Nat Methods*, **7**, 248–249. https://doi.org/10.1038/nmeth0410-248

[8] Li, M.X., Kwan, J.S., Bao, S.Y., Yang, W., Ho, S.L., *et al*. (2013) Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genet*, **9**, e1003143. https://doi.org/10.1371/journal.pgen.1003143

[9] Gnad, F., Baucom, A., Mukhyala, K., Manning, G. and Zhang, Z. (2013) Assessment of Computational Methods for Predicting the Effects of Missense Mutations in Human Cancers. *BMC Genomics,* **14**, S7.

[10] Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res,* **39**, e118. https://doi.org/10.1093/nar/gkr407

[11] Dehouck, Y., Kwasigroch, J.M., Rooman, M. and Gilis, D. (2013) BeAtMuSiC: prediction of Changes in Protein-Protein Binding Affinity on Mutations. *Nucleic Acids Res*, **41**, W333–339. https://doi.org/10.1093/nar/gkt450

[12] Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm. *Nature Protocols*, **4**, 1073–1081. https://doi.org/10.1038/nprot.2009.86

[13] Dehouck, Y., Kwasigroch, J.M., Gilis, D. and Rooman, M. (2011) PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. *BMC Bioinformatics*, **12**,151. https://doi.org/10.1186/1471-2105-12-151

[14] Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., *et al*. (2010) A Method and Server for Predicting Damaging Missense Mutations. *Nat Methods*, **7**, 248–249. https://doi.org/10.1038/nmeth0410-248

[15] Song, Y. X., Zhou, X., Wang, Z., Gao, P., Li, A.L., *et al*. (2012) The Association between Individual SNPs or Haplotypes of Matrix Metalloproteinase 1 and Gastric Cancer Susceptibility. *Progression and Prognosis*, **7**, e 38002.

[16] Hamosh, A., Scott, A.F. Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders. *Nucleic Acids Research*, **33**, D514–D517.

https://doi.org/10.1093/nar/gki033
[17] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., *et al.* (2011) The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored. *Nucleic Acids Res*, **39**, D561–8. https://doi.org/10.1093/nar/gkq973
[18] Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000. *Nucleic Acids Res*, **28**, 45–48. https://doi.org/10.1093/nar/28.1.45
[19] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005) The FoldX Web Server: An Online Force Field. *Nucl. Acids Res,* **33** W382–8. https://doi.org/10.1093/nar/gki387
[20] Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M. and Gaunt, T.R. (2013) Predicting the Functional Consequences of Cancer-Associated Amino Acid Substitutions. *Bioinformatics,* **29**,1504-1510. https://doi.org/10.1093/bioinformatics/btt182
[21] Thomas, P.D., Kejariwal, A., Guo, N., Mi, H.Y. and Campbell, M.J., Muruganujan, A. and Lazareva-Ulitsky, B. (2006) Applications for Protein Sequence-Function Evolution Data: mRNA/protein Expression Analysis and Coding SNP Scoring Tools. *Nucl. Acids Res*, **34**, W645-W650. https://doi.org/10.1093/nar/gkl229
[22] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5–32. https://doi.org/10.1023/A:1010933404324
[23] Cortes, C. and Vapnik, V. (1995) Support Vector Machine. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/A:1022627411411
[24] Stein, A. and Aloy, P. (2010) Novel Peptide-Mediated Interactions Derived from High-Resolution 3-dimensional Structures. *PloSComput. Biol*, **6**, e1000789. https://doi.org/10.1371/journal.pcbi.1000789
[25] Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D. (2005) The International Hap Map Project Website. *Genome Res*, **15**, 1592-1593. https://doi.org/10.1101/gr.4413105
[26] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The Human genome browser at UCSC. *Genome Res*, **12**, 996-1006. https://doi.org/10.1101/gr.229102
[27] McVean, *et al*. (2012) An Integrated Map of Genetic Variation from 1092 Human Genomes. *Nature*, **491**, 56-65. https://doi.org/10.1038/nature11632
[28] Hanley, J. and McNeil, B. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Supplementary Data

S1: Feature name and description about each feature.

No.	Feature Name	Description
1	Fatthm Score	Fatthm score determining the cancerous nature of SNP calculated from fatthm tool
2	SubPSEC	Substitution Position-Specific Evolutionary Conservation (subPSEC) Score
3	Pdeleterious	Probability that a given variant will cause a deleterious effect on protein function calculated by Panther tool
5	Total Energy	Total energy difference of wild and mutant type
6	Backbone HBond	The contribution of backbone Hbonds
7	Sidechain Hbond	The contribution of sidechain-sidechain and sidechain-backbone Hbonds
8	Vander Waals	Contribution of the Vander Waals

9	Electrostatics	Electrostatic interactions
10	Solvation Polar	Penalization for burying polar groups
11	Solvation Hydrophobic	Contribution of hydrophobic groups
12	Vander Waals clashes	Energy penalization due to Vander Waals' clashes (interresidue)
13	Entropy side chain	Entropy cost of fixing the side chain
14	Entropy main chain	Entropy cost of fixing the main chain
15	Torsional clash	Vander Waals' torsional clashes (intraresidue)
16	Backbone clash	Backbone-backbone Vander Waals.
17	Helix dipole	Electrostatic contribution of the helix dipole
18	Disulfide	Contribution of disulfide bonds
19	Electrostatic kon	Electrostatic interaction between molecules in the pre-complex
20	Partial covalent bonds	Interactions with bound metals
21	Energy ionisation	Contribution of ionisation energy