IN SILICO SCREENING OF DELETERIOUS SNPs OF GENE *U2AF1* FOR THEIR ROLE IN LUNG CANCER

Enrollment Number: 151509

Name of Student: Shweta Parmar

Name of Supervisor: Dr. Tiratha Raj Singh

MAY 2019

Thesis submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNGHAT, SOLAN

Abstract

Lung cancer is another most common cancer worldwide. Various genes are acknowledged which are responsible for imparting the vulnerability to lung cancer. Amongst those genes, *U2AF1* is identified as the major risk factor. *U2AF1* is known to cause missense mutations in the several types of cancer together with lung adenocarcinoma (LUAD) and myeloid leukemia (AML). It affects the biological pathway such as DNA methylation, X chromosome inactivation and alters the selected 3` splice site motif. In this study, we present in silico screening and molecular dynamic simulation of lung cancer associated deleterious non-synonymous single nucleotide polymorphisms in *U2AF1*. We have identified three deleterious coding non-synonymous single nucleotide polymorphisms rs371246226 (Q157R), rs17850009 (G94R), rs371769427 (S34F) in *U2AF1* using computational tools SIFT, Polyphen2, PANTHER, SNPs&GO, PhD-SNP, MutPred, SNAP, PROVEAN for the sequence based analysis and tools SNPs&GO^{3D}, PANTHER for the structure based analysis. We have performed molecular dynamics simulations to predict the structural effects of these *U2AF1* mutations comparative to the wild-type protein. Results from our simulations demonstrated a comprehensive effect of the mutations that could be able to provide foresight for therapeutic methods in lung cancer.

CHAPTER 1

INTRODUCTION

Lung Cancer is the most fatal form of a cancer and is the primary cause of cancer mortality. After the breast cancer in women and prostate cancer in men lung cancer is the second most common cancer worldwide. In 2010, estimated cases for lung cancer were approximately 116,000 in men and 105,000 in women, with approximately 160,000 deaths. Lung cancer accounted for 31% and 27% of overall deaths in men and women [1]. In 2012, estimated cases for lung cancer were 1,800,000, 58% of which occurred in less developed regions. With 1,590,000 (19.4%) of cancer deaths worldwide. In 2017, American Cancer Society estimated 222,500 cases of lung cancer and approximately 155,000 individuals died as a result of the disease [2].

Lung cancer typically refers to the malignant bronchogenic epithelial tumors of the lung, namely, squamous cell carcinoma, large cell carcinoma, adenocarcinoma and small cell carcinoma. It can spread to any organ like liver, brain, lymph glands, spinal cord, bones and adrenal glands. Symptoms of this metastatic disease includes, weight loss but also fatigue, poor appetite, weakness, and novel symptoms shown by any organ including musculoskeletal pain, neurologic change, and abdominal discomfort[1].

It is believed that smoking and lung cancer are linked but there is no certain acknowledgement that why lung cancer evolved in massive smokers and not in others. It is assumed that due to exposure to carcinogens genetic factors increase the risk of lung cancer in some individuals. Genetic factors include specific enzymes which metabolizes the products of cigarette to dominant carcinogens, as one of the enzyme aryl hydrocarbon hydroxylase is induced by smoking and it converts hydrocarbons to carcinogenic metabolites. Yet there are unidentified genetic factors which may affect the sensitivity to carcinogens or may include the tumor suppressor genes activity [3].

It is a complex disease as it not only affects biochemical level that is genes or proteins but it also affect at the tissue, organism and population levels. That is why there are many early detection biomarkers which include tissue based biomarkers and biofluid based biomarkers such as sputum, exhaled breath, and blood and airway epithelium. But detection by one single biomarker is a difficult task due to the heterogeneity of lung cancer as these often flap with other cancers and inflammatory conditions [4].

By examining the cases of lung adenocarcinoma and squamous cell carcinoma, TCGA (The Cancer Genome Atlas) showed a set of fifty three genes significantly mutated and associated with lung cancer risk [5]. Among the fifty three SMGs conferring high lung cancer risk *U2AF1* (U2 small nuclear RNA auxiliary factor 1) also known as U2AF35 is known to be mutated in several types of cancer together with lung adenocarcinoma (LUAD) and acute myeloid leukemia (AML).

U2AF1 (U2 auxiliary factor protein) is the small subunit 35kDa of U2 snRNP auxiliary factor (U2AF) with 240 residues. It belongs to the splicing factor SR family and it recognizes the AG splice acceptor dinucleotide at the 3` end of introns. *U2AF1* plays role in pre-mRNA processing that is RNA splicing and it forms a heterodimer with U2AF65 the large subunit (U2AF2) of U2AF. The smaller subunit attach 3` AG splice acceptor dinucleotide of the pre-mRNA target intron and the larger subunit attach the adjoining poly pyrimidine region [6].

U2AF35 plays a vital role in each constitutive splicing and enhancer-dependent splicing by mediating protein-protein interactions and protein-RNA interactions as it acts as a mediator for enhancer-dependent splicing and is required for constitutive splicing. It directly intervene the interactions between the U2AF65 and proteins bound to the enhancer, resulting in recruiting U2AF65 to the adjacent intron by acting as a bridge between the enhancer complex and the U2AF2 [7,8] (shown in Figure 1).

This protein coding gene has (i) two zinc-finger regions, C3H1-type1 (12-40) and C3H1-type2 (149-176), (ii) SR-rich domain at the C-terminal operates the interaction with SR proteins and the splicing regulators namely TRA and TRA2 (iii) domain at the N-terminal is involved in the formation of the U2AF1/U2AF2 heterodimer [8].

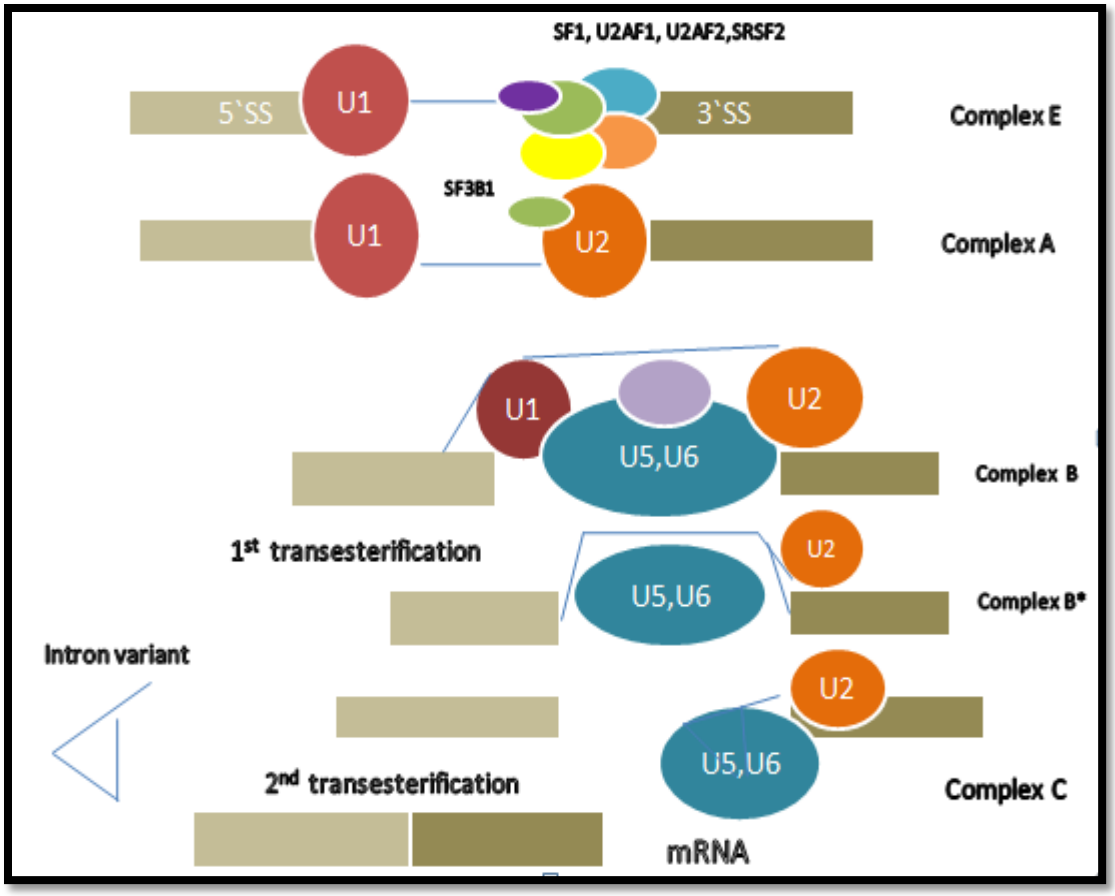


Figure1 Several distinct complexes during RNA splicing. Binding of U1snRNP and U2AF1/2 to the transcribed pre-mRNA forms the first complex. By the formation of complex E recruitment of U2snRNP activated and forms complex A. Complex B is formed as U4/U6.U5 tri-SNP complex joins. Complex B is catalytically activated i.e. complex B^{*}, after the release of U1/U4 snRNPs. Conformational change occurs and resulting in the formation of complex C. Eventually complex C catalyzes the second esterification reaction and removes the intron as intron lariat. snRNP: small nuclear ribonucleoprotein. U2AF: U2 small nuclear RNA auxiliary factor.

It is indentified that *U2AF1* mutations are missense, resulting in the substitution of a single amino acid with another that codes for a different amino acid. It creates differential splicing of hundreds of genes which affects the biological pathways such as DNA methylation, X chromosome inactivation, the DNA damage response, and apoptosis. It also changes the selected 3` splice site motif and affect the first and second zinc fingers [9].

Single nucleotide polymorphisms (SNPs) are genetic variations within the human genome. DNA mismatch pair, cell cycle regulation and immunity are regulated by the SNPs within genes corresponding gene status to cancer. SNPs can be located in distinct regions such as promoters, exons, introns and in 5`- and 3`- untranslated regions (UTRs). Variations in the gene expression and their consequences to cancer depend on the location of SNPs. As (i) SNPs present in the promoter region changes the function of promoter that is DNA methylation, histone modification and transcription factor binding activity which eventually affect the gene expression. (ii) exon region SNPs inhibit the transcription and translation activity. (iii) SNPs of intron regions effect splice alternatives of transcripts and either damage or encourage the activity of long non-coding RNAs (lncRNAs). (iv) SNPs of 5`-UTRs modify the translation activity and SNPs of 3`-UTRs micro RNA binding activity [10].

Predicting how these single nucleotide polymorphisms (SNPs) influence the function of proteins is a crucial domain of a research. To understand the molecular basis of disease in genetic studies potential SNP selection is important and it is possible by efficiently identifying such SNPs. Non synonymous coding SNPs (nSNPs) are the SNPs which cause alteration in the amino acids of protein sequence of gene, these SNPs are located in the coding regions and have enormous effect on the phenotype [11].

Our project work has focused on the non synonymous coding SNPs located in the coding region of *U2AF1* gene which have impact on the lung cancer phenotype. By exploring different computational algorithm tools SIFT, Polyphen-2, Provean, SNP&GO, PhD-SNP, PANTHER, MutPred, SNAP for classifying the deleterious lung cancer associated nSNPs.

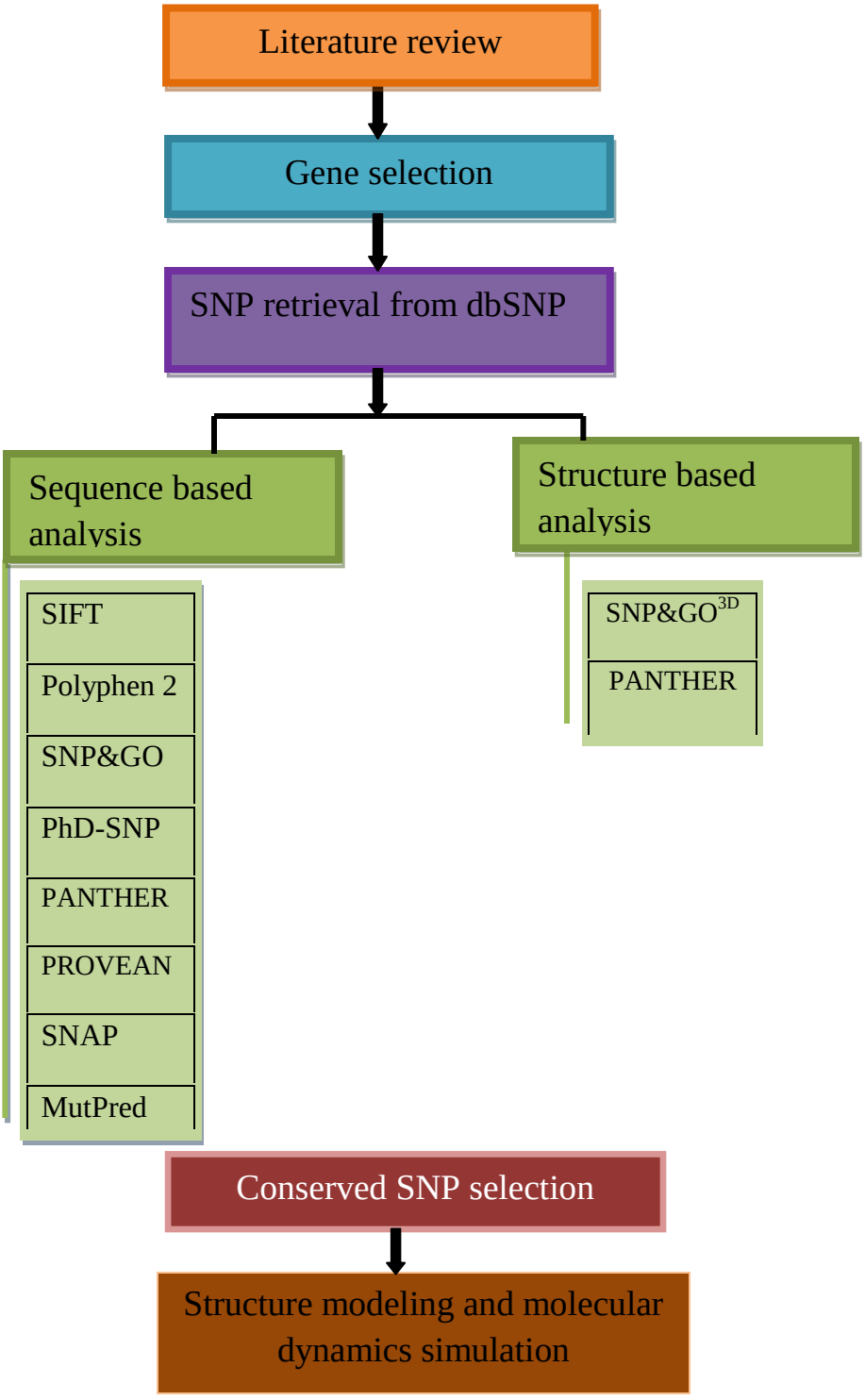
Molecular dynamics simulation (MDS) is an essential tool for analyzing the impact of mutations on the protein structure. Molecular dynamics simulation have been performed of WT and three selected mutants (S34F, G94R, Q157R) to check (i) the effect of mutants on the conformation in the functional regions of native protein (ii) the deviation of the mutant structures from the native (iii)whether the mutants are causing alteration in the flexibility of native protein.

The objective of our project work is to identify the lung cancer associated deleterious coding non synonymous SNPs and to further predicts the structure level behavior after mutation through molecular dynamic simulation.

CHAPTER-2

MATERIALS AND METHODS

Protocol followed is shown below:



2.1 Data Retrieval

This chapter includes the materials and methods of the project work. In this chapter all the computational and explication steps are thoroughly discussed. Selection of the gene *U2AF1* and understanding its relation to lung cancer was done by reviewing various literatures [5, 6, 9]. *U2AF1* SNPs were retrieved from dbSNP database (https://www.ncbi.nlm.nih.gov/snp/u2af1. accessed: 16- Oct- 2018) for computational analysis [12].

2.2 Prediction of deleterious coding non synonymous SNPs (nSNPs)

We used **sequence based analysis** tools SIFT, Polyphen-2, Provean, SNP&GO, phD-SNP, PANTHER, MutPred, SNAP to filter out the deleterious coding SNPs from other SNPs for *U2AF1*.

- **SIFT:** 'Sorting Tolerant from Intolerant' (SIFT) (https://sift.bii.a-star.edu.sg/www/SIFT_dbSNP.html. accessed: 16- Oct- 2018) predicts if the substitution of an amino acid in a protein will have an effect on its function or not by using sequence homology based approach [13]. When a protein sequence is given, SIFT select the associated proteins and the alignment are acquired with the query sequence. Score is assigned to every residue. Scores ranges from 0-1, result ranging from 0-0.05 are evaluated as deleterious substitutions and 0.05-1 scores are evaluated as tolerant or neutral amino acid substitutions [14]. Our query dbSNP ids were submitted as an input to SIFT server.
- **Polyphen-2:** 'Polymorphism phenotyping v2' (http://genetics.bwh.harvard.edu/pph2/. accessed: 17- Oct- 2018). uses Naïve Bayes classifier model to identify the functional effect of an amino acid substitution. Using two datasets, prediction models of this web based server were tested and trained accordingly.

  First is HumanDiv and second is HumVar. HumanDiv dataset is assembled from every harmful allele exist in the UniProtKB database, whose impacts are known to cause human mendelian diseases due to impact on the molecular function, in addition to variation between human proteins and the mammalian homologs at a close range with them, are supposed to be non-damaging. On the other hand, HumanVar dataset contains all the diseases from UnitProtKB database which cause mutations along with the some

common human non synonymous SNPs which do not cause diseases, these nsSNPs are considered as non-damaging. Depending on the false positive rate (FPR) thresholds, mutation is evaluated as possibly damaging, probably damaging and benign [15].

- **PROVEAN:** 'PROtein Variation Effect Analyzer' (http://provean.jcvi.org/index.php. accessed: 17- Oct- 2018) predicts the impact on the function of protein due to alteration in the protein sequence. To generate the prediction, query sequence is given sequences related to the query are collected from the BLAST and clustering of hits is performed on the basis of seventy five percent global sequence identity. A supporting sequence set is formed of about top thirty clusters most related to query. Delta alignment score is evaluated for each supporting sequences, this score is based on the alternative query sequence relative to sequence homologs obtained from NCBI protein database using BLAST. Default threshold for the classification is -2.5, score equal or below this value is considered to have deleterious effect and the score above threshold is predicted to have neutral effect [16].
- **SNPs&GO:** 'Single Nucleotide Polymorphism Database & Gene Ontology' (https://snps-and-go.biocomp.unibo.it/snps-and-go/index.html. accessed: 19- Oct- 2018). is a web based server which identifies the mutations in proteins associated with human disease with functional interpretation. This method is based on the support vector machines (SVM) to anticipate the mutations from the protein sequence. For scoring, measurement of the quality of binary classification is based on the Matthews correlation coefficient equals to 0.63 and accuracy of eighty two percent. This SVM based system exerts information from the gene ontology observation to predict if the mutation can be classified as disease related or not. Using over 33000 mutations sets and cross validation procedure sets wherein similar proteins were held to the same dataset to calculate the LGO score resulting from the GO data base , this server was trained and tested [17].
- **PhD-SNP:** 'Predictor of human Deleterious Single Nucleotide Polymorphisms' (http://snps.biofold.org/phd-snp/phd-snp.html. accessed: 21- Oct- 2018) is a classifier which predicts deleterious SNPs. It is based on a support vector machine classifier. For a given query sequence, this classifier categorized the mutations into binary categories: Disease and Neutral that is whether the mutation is disease related or it is neutral polymorphism, 0.5 is a threshold value set for the conclusion. Output of PhD-SNP

classifier consists of probability of correct predictions and the RI (Reliability Index) value which is estimated form the SVM output [18].

- **PANTHER:** 'Protein Analysis THrough Evolutionary Relationships' (http://www.pantherdb.org. accessed: 22- Oct- 2018) is a classification approach to classify the proteins. PANTHER basically have two sections, one is PANTHER library consists of protein family which is represented as multiple sequence alignment (MSA), a family tree and the Hidden Markov Model (HMM) and another is PANTHER index is composed of an ontology which is used to study and compile the molecular functions and biological processes of the families. To predict the deleterious or missense SNPs according to their impact on the protein function position specific score of the HMMs family are used [19].
- **MutPred:** 'MUTation PREDiction' (http://www.mutdb.org. accessed: 25- Oct- 2018) is a prediction method for variations in a coding region which alter the processing of pre-mRNA. This dataset is based on two training sets: first is Splice Affecting Variant (SAV) and the second is Splice Neutral Variant (SNV). These two data sets have set a threshold value of 0.6 to predict the alteration which damages the splicing process. Output of the MutPred is in the form of score which is basically the probability that the amino acid substitution is deleterious. Output score of variants greater than the threshold value is considered as a SAV and less than threshold value is considered as a SNV [20].
- **SNAP:** 'Screening for Non Acceptable Polymorphisms' (https://rostlab.org/services/snap2web/. accessed: 27- Oct- 2018) is a method build on neural-network (Artificial Neural Network) classifier to identify the consequences on protein function due to the substitution of non synonymous SNPs. ANN is not an algorithm and works as a system for many algorithms of machine learning to process and the input data. It is build of connected nodes collected together. Scoring of the SNAP ranges from -100 to +100. -100 suggests prediction to be strong neutral that is mutation did not change the function of native protein whereas +100 is considered as a prediction of strong effect of the mutation on the function of native protein [21].

Structure based analysis is done by using the software SNP&GO^{3d} and PANTHER to investigate the effects of mutations at the structural level.

- **SNP&GO^{3D}:** This approach is based on the SVM classifier. Input is given as a 3D structure of a protein, information of mutations and functional information as gene ontology terms. To predict the results, information of the input mutations and protein is analyzed by the classifier. A vector is formed of 20 elements in which substitutions are encoded from the WT residue to the mutants, and a second vector of 20 elements which encodes for the structural environment of mutated residue in 3D structure. Relative solvent accessible area (RSA) of the residues that are mutated is computed by using the DSSP algorithm. Profile features of an input protein are abstracted from the BLAST. A frequency of the WT residues and the mutated residues at a particular position is assessed [23].
- **PANTHER:** This database uses the 3D structure of protein to predict the disease related mutations. It goes for machine learning application which is structure based that is SVM. It is trained from a several protein chains having set of disease related mutations and neutral polymorphisms [22]. PANTHER algorithm to calculate the frequency of the WT residue and the residues that are mutated, to identify the likelihood of the mutations that is deleterious [23].

2.3 Structure Modeling

Structure modeling was done by SWISS model [24] using homology modeling method. The template is taken (PDB id: 4yh8) from the schizosacchromyces pombe. The structure is crystallized in 1.7 Armstrong resolutions and U2AF2 domain is also bound with the structure. In our structure we have also taken U2AF2 fragment for predicting that how mutation is inducing the structural changes and how it alters the binding of *U2AF1* and U2AF2 gene. After that the mutation is constructed by using CHIMERA 1.13.2 (https://www.cgl.ucsf.edu/chimera/). Then the mutant structures S34F, G94R, Q157R with the wild-type (WT) were implied for Molecular Dynamic Simulation (MDS).

2.4 Molecular Dynamic Simulation

To examine the effect of the mutations on *U2AF1* at the structural level molecular dynamics was performed. Analysis was performed by using GROMACS 4.5.3 (http://www.gromacs.org/) at physiological temperature 37^0C and neutral pH. To investigate the mutation induced effect on a structural level we have carried out 100ns (nano second) MDS. It reveals the structural dynamics of the WT as well as mutants. All the structures were solvated in a triclinic box. The protein topology was generated by using AMBERff99SB force field and then the systems were neutralized by the addition of two Cl^- ions. After that systems were implied for the energy minimization by using steepest descent and conjugant gradient method. Consequently the 1ns NVT and NPT simulation were performed for maintaining the temperature, volume and pressure at 300K. The coordinates were saved in a time step of 2 fs. After each five steps, record of non bonded pair was updated according to the position restraints for heavy atoms and LINCS constraints for all bonds [25]. In order to conserve the temperature stability Berendsen thermostat was used [26]. Computation of electrostatic interactions was done by particle mesh Ewald summation method [27]. Finally, all four (WT, S34F, Q157R, G94R) were implied for 100ns MDS.

Analysis of Molecular dynamics simulations

By operating the in-built functions of GROMACS package g_rms, g_rmsf, g_sas, g_gyrate, g_hbond, examination of structural alterations in native and mutated structures such as root mean- square deviation (RMSD), root mean-square fluctuation (RMSF), solvent-accessible surface area (SASA), radius of gyration (Rg), and protein-solvent intermolecular hydrogen bonds (H-bonds) were evaluated. The plot was generated by using ORIGIN and the trajectories were analyzed by using CHIMERA.

CHAPTER 3

RESULT AND DISCUSSION

Prediction of SIFT server

Three dbSNPs were predicted deleterious having score less than 0.05 approximately equals to 0 and rest of the dbSNPs were predicted as tolerated mutations with score greater than 0.05. Nucleotide change of: C/T was shown by three nSNPs, T/C was exhibited by one nSNP, C/A and G/C was exhibited by one and two nSNPs exhibited change from G/A. Alteration of the amino acids on the other hand varies as: two exhibited the alteration at the Arginine region, three nSNPs exhibited the alteration at the Glycine region, two exhibited the alteration at the Serine region and one showed a alteration at the Glutamine region (Table 1).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	SIFT score	PREDICTION
rs371246226	T/C	Q157R	ENSP00000291552	0.001	Deleterious
rs371769427	G/A	S34F	ENSP00000291552	0	Deleterious
rs17850009	C/T	G94R	ENSP00000291552	0.004	Deleterious
rs140932020	C/T	G222S	ENSP00000291552	0.476	Tolerated
rs200044775	G/A	S231L	ENSP00000291552	0.443	Tolerated
rs202230168	C/A	G210V	ENSP00000291552	0.083	Tolerated
rs371518817	C/T	R205Q	ENSP00000291552	0.078	Tolerated
rs375393848	G/C	R119G	ENSP00000291552	0.321	Tolerated

Table1: Prediction scores by SIFT server.

Prediction of Polyphen2 server

Polyphen-2 estimated if the alteration is benign, possibly damaging or probably damaging. It uses HumDiv and HumVar Bayesian stochastic templates for the prediction. For the diagnosis of mendelian disease HumVar model should be used because it can recognize the variations having harmful impact from the variations having no harmful impact HumDiv on the other hand is appropriate for the prediction of the mutants moderately deleterious genes are considered as damaging [15]. Four nSNPs were predicted as probably damaging by both HumDiv and HumVar. Four were predicted as possibly damaging by only HumDiv. Two were identified as benign by HumVar and the same nSNPs were identified as possibly damaging by HumDiv (Table 2).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	HumDiv Score	HumVar Score	PREDICTION (HumDiv/ HumVar)
rs371246226	T/C	Q157R	ENSP00000291552	0.997	0.971	Probably damaging
rs371769427	G/A	S34F	ENSP00000291552	1	0.997	Probably damaging
rs17850009	C/T	G94R	ENSP00000291552	1	1	Probably damaging
rs140932020	C/T	G222S	ENSP00000291552	0.996	0.97	Probably damaging
rs200044775	G/A	S231L	ENSP00000291552	0.934	0.706	Possibly damaging
rs202230168	C/A	G210V	ENSP00000291552	0.531	0.312	Possibly damaging / Benign
rs371518817	C/T	R205Q	ENSP00000291552	0.934	0.614	Possibly damaging
rs375393848	G/C	R119G	ENSP00000291552	0.588	0.111	Possibly damaging / Benign

Table2: Prediction scores of HumDiv and HumVar by Polyphen-2 server

Prediction of PhD-SNP

PhD-SNP estimated if the alteration is disease related or is neutral. Reliability index which is the output of SVM classifier shows whether the mutation is reliable or not and the probability indicates the accuracy of the prediction. This tool has predicted two nSNPs causing alteration at the serine region and at the glycine region as disease related mutations that are not reliable and the rest of the SNPs were predicted as neutral polymorphism (Table 3).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	Reliability Index	Probability	PhD-SNP Prediction
rs371246226	T/C	Q157R	ENSP00000291552	0	0.498	Neutral
rs371769427	G/A	S34F	ENSP00000291552	5	0.737	Disease
rs17850009	C/T	G94R	ENSP00000291552	5	0.759	Disease
rs140932020	C/T	G222S	ENSP00000291552	5	0.267	Neutral
rs200044775	G/A	S231L	ENSP00000291552	6	0.210	Neutral
rs202230168	C/A	G210V	ENSP00000291552	3	0.327	Neutral
rs371518817	C/T	R205Q	ENSP00000291552	7	0.143	Neutral
rs375393848	G/C	R119G	ENSP00000291552	1	0.428	Neutral

Table3: Prediction scores by PhD-SNP tool.

Prediction of SNP&GO

SNP&GO predicts the alteration and categories the output into two categories, one is disease related and another is neutral alterations. Range of RI goes from 0 to 1, and the RI score represents how much reliable the alteration is. As score closer to 0 indicates that the mutation is not reliable and predicts the mutation as disease related polymorphism and score closer to 10 indicates that the mutation is reliable and predicts the mutation as neutral polymorphism. Result predicted for the given dbSNPs by the this tool shows, two nSNPs causing mutations related to disease having RI less than 5 and rest of the nSNPs are neutral polymorphisms having RI more than or equal to 5 (Table 4).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	Reliability Index	Probability	SNP&GO Prediction
rs371246226	T/C	Q157R	ENSP00000291552	4	0.315	Neutral
rs371769427	G/A	S34F	ENSP00000291552	3	0.64	Disease
rs17850009	C/T	G94R	ENSP00000291552	3	0.63	Disease
rs140932020	C/T	G222S	ENSP00000291552	8	0.106	Neutral
rs200044775	G/A	S231L	ENSP00000291552	9	0.06	Neutral
rs202230168	C/A	G210V	ENSP00000291552	6	0.203	Neutral
rs371518817	C/T	R205Q	ENSP00000291552	9	0.061	Neutral
rs375393848	G/C	R119G	ENSP00000291552	5	0.269	Neutral

Table4: Prediction scores by SNP&GO tool.

Prediction of PANTHER

Prediction of this tool is based on the Hidden Markov Model. PANTHER predicts five nSNPs as disease related mutations out which two nSNPs showed amino acid changes at the serine region and other two showed the amino acid changes at the glycine region. One nSNPs showed the amino acid change at the glutamine region. Rest of the nSNPs was predicted as neutral mutations (Table 5).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	Reliability Index	Probability	PANTHER Prediction
rs371246226	T/C	Q157R	ENSP00000291552	1	0.565	Disease
rs371769427	G/A	S34F	ENSP00000291552	9	0.969	Disease
rs17850009	C/T	G94R	ENSP00000291552	8	0.923	Disease
rs140932020	C/T	G222S	ENSP00000291552	6	0.205	Neutral
rs200044775	G/A	S231L	ENSP00000291552	5	0.744	Disease
rs202230168	C/A	G210V	ENSP00000291552	1	0.568	Disease
rs371518817	C/T	R205Q	ENSP00000291552	1	0.458	Neutral
rs375393848	G/C	R119G	ENSP00000291552	0	0.485	Neutral

Table5: Prediction scores by PANTHER.

Prediction of PROVEAN

PROVEAN predicts the functional impact of alteration at the position of amino acids in the protein sequence. Cut off value of the prediction score is -2.5. PROVEAN categorizes the prediction into two parts one is deleterious SNPs and another is neutral SNPs. This server has predicted five nSNPs as deleterious SNPs having the prediction score less than or equal to -2.5 and three SNPs as neutral SNPs having the prediction score greater than -2.5.

Deleterious SNPs have mutated the amino acids at the two serine regions, at the glutamine region, at the glycine and arginine regions (Table 6).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	PROVEAN Score	PREDICTION
rs371246226	T/C	Q157R	ENSP00000291552	-3.69	Deleterious
rs371769427	G/A	S34F	ENSP00000291552	-5.64	Deleterious
rs17850009	C/T	G94R	ENSP00000291552	-7.79	Deleterious
rs140932020	C/T	G222S	ENSP00000291552	-1.16	Neutral
rs200044775	G/A	S231L	ENSP00000291552	-2.54	Deleterious
rs202230168	C/A	G210V	ENSP00000291552	-2.35	Neutral
rs371518817	C/T	R205Q	ENSP00000291552	-1.54	Neutral
rs375393848	G/C	R119G	ENSP00000291552	-3.51	Deleterious

Table6: Prediction scores by PROVEAN.

Prediction of SNAP

SNAP is based on the ANN and it screens out the harmful polymorphisms from the protein sequence. Screening of the given dbSNPs is based on the scoring system of the network. Score ranges from -100 to +100. Prediction is based on the functional effect of the mutations. In the extracted output it is shown that nSNPs which have score closer to the 100 and more than 50 are predicted as the polymorphisms having strong effect on the function of protein. Whereas nSNPs has score below 50 are categorized as having weak effect and score below 0 as no effect on the function of the protein that is having no drastic impact (Table 7).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	SNAP Score	PREDICTION
rs371246226	T/C	Q157R	ENSP00000291552	63	Strong effect
rs371769427	G/A	S34F	ENSP00000291552	66	Strong effect
rs17850009	C/T	G94R	ENSP00000291552	79	Strong effect
rs140932020	C/T	G222S	ENSP00000291552	-20	No effect
rs200044775	G/A	S231L	ENSP00000291552	30	Weak effect
rs202230168	C/A	G210V	ENSP00000291552	-31	No effect
rs371518817	C/T	R205Q	ENSP00000291552	31	No effect
rs375393848	G/C	R119G	ENSP00000291552	49	No effect

Table7: Prediction by SNAP tool.

Prediction of MutPred

MutPred output consists of the probability scores and the property scores. Probability score predicts that if the amino acid alteration is disease associated or not and property score is the P-value that defines that structural and functional properties of the protein are affected. Hypotheses are made on the basis of general score and property score. A hypothesis basically is the definite combination of high values of general scores and low values of property scores. There are three types of hypotheses given as an output by this tool: actionable hypotheses, confident hypotheses, very confident hypotheses. Probability score greater than 0.5 is predicted as actionable hypotheses that is mutation is harmful, score greater than 0.75 is predicted as confident and very confident hypotheses that is mutation is highly harmful (Table 8).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	MutPred Score	PREDICTION
rs371246226	T/C	Q157R	ENSP00000291552	0.504	Harmful
rs371769427	G/A	S34F	ENSP00000291552	0.831	Highly harmful
rs17850009	C/T	G94R	ENSP00000291552	0.556	Harmful
rs140932020	C/T	G222S	ENSP00000291552	0.324	Not harmful
rs200044775	G/A	S231L	ENSP00000291552	0.296	Not harmful
rs202230168	C/A	G210V	ENSP00000291552	0.330	Not harmful
rs371518817	C/T	R205Q	ENSP00000291552	0.227	Not harmful
rs375393848	G/C	R119G	ENSP00000291552	0.317	Not harmful

Table8: Prediction by MutPred tool.

Prediction of SNP&GO^{3D} and PANTHER

These structure based analysis tools predict only disease associated mutations in the protein structure using gene ontology terms. Nucleotide change of C/T and G/C shown in the output. Amino acid mutations at the glycine region and at the arginine region are predicted by these two servers (Table 9).

DbSNPID	Nucleotide change	Amino Acid change	Protein ID	SNP&GO^{3D}	PANTHER
rs17850009	C/T	G94R	ENSP00000291552	Disease	Disease
rs375393848	G/C	R119G	ENSP00000291552	Neutral	Neutral

Table9: Prediction by SNP&GO^{3D} and PANTHER.

Molecular dynamics simulation analysis

Conclusion from the computation of root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg), solvent accessible surface area (SASA), protein-solvent intermolecular hydrogen-bonds (H-bonds) for WT, S34F, G94R, Q157R presented in the Table 10.

	Initial WT	S34F	G94R	Q157R
RMSD (nm)	0.31	0.41	0.25	0.38
RMSF (nm)	0.16	0.17	0.15	0.21
Rg (nm)	1.71	1.73	1.72	1.77
SASA	124.16	123.59	123.66	127.54
H-bonds	146.43	147.37	150.24	144.10

Table10: Time averaged structural properties estimated for WT, S34F, G94R and Q157R. RMSD: root mean square deviation, RMSF: root mean square fluctuation, Rg: radius of gyration, SASA: solvent accessible surface area, H-bonds: intermolecular hydrogen bonding.

RMSD analysis:

To examine the effect of mutants, we have studied the RMSD values. As RMSD is a foremost region to assess the protein system. The computed RMSDs of the atoms in WT, S34F, G94R, Q157R with respect to the initial structure during the 100 ns MDS as a function of time is plotted in the Figure 2. Throughout the RMSD of Q157R showed an increase in the initial 15ns with a sudden sharp increase around 20ns followed by the equilibrium around 60ns and a sudden increase after 80ns. Although, S34F and G94R showed a different trend of RMSD. As S34F showed a sudden increase from the initial ns to 100ns with respect to WT. On the other hand G94R showed a less variations in the initial ns followed by equilibrium around 0 to 50ns and a sudden decrease after 60ns followed by equilibrium again after 82ns. A comparison of average RMSD values exhibit the following order of structural deviation (Table 10): S34F > Q157R >

WT > G94R. This specifies that a serious change was observed in the S34F compared to the other variants.

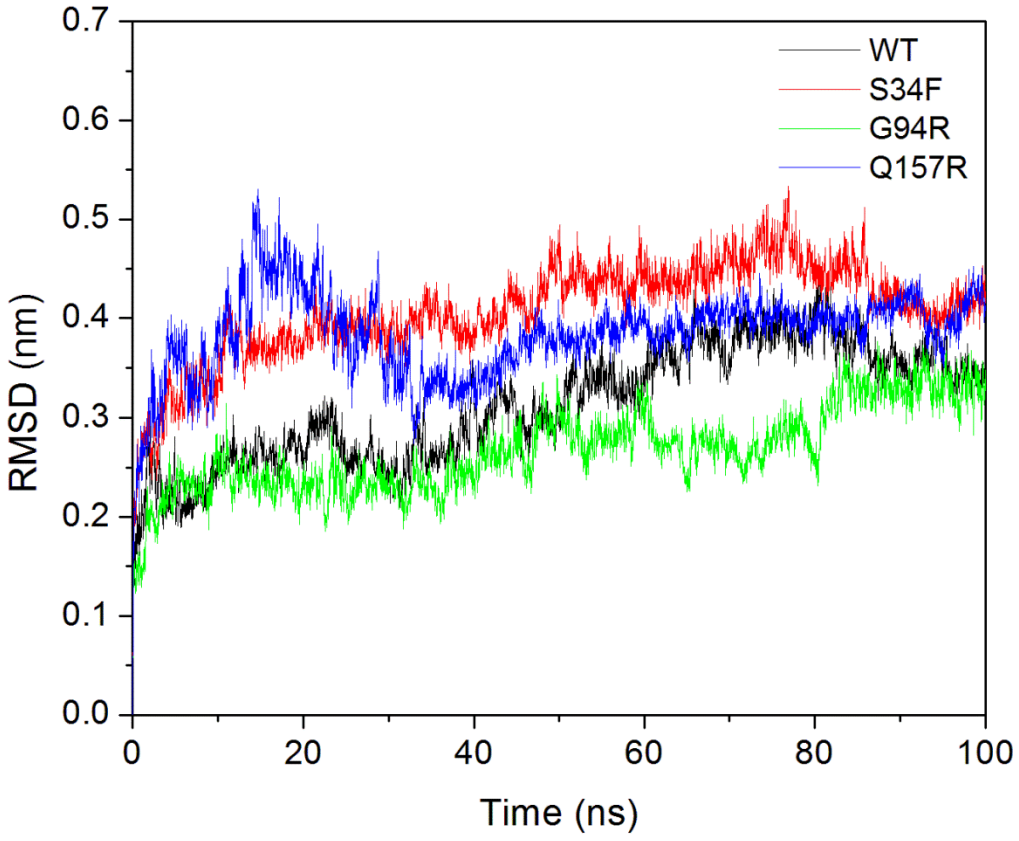


Figure 2: Figure shown represents WT, S34F, Q157R, G94R RMSD plot as a function of time.

RMSF analysis:

So to figure out by what means mutants influence the potent nature of the residues also to analyze what roots the conformational shifts obtained in the evaluation of RMSD, root mean square fluctuation of WT and mutated AA residues were computed and plotted in the Figure 3. Except G94R, rest of the mutants had higher average RMSFs that WT simulation (Table 10): Q157R > S34F > WT > G94R. Mutants showed an increase and decrease of the flexibility. Above 50 percent of the residues hold RMSF > 0.1nm in the WT, signifying great rank of variation. Whereas, each mutants has shown a major proportion of residues with RMSF value greater than 0.1nm. As S34F has shown a higher RMSF value 0.76nm, Q157R showed a RMSF

0.9nm and G94R showed a higher RMSF value 067nm. Among these mutants, Q157R have higher proportion of residues with RMSF 0.98nm, consequently remarking a higher impact on the whole flexibility of the native protein.

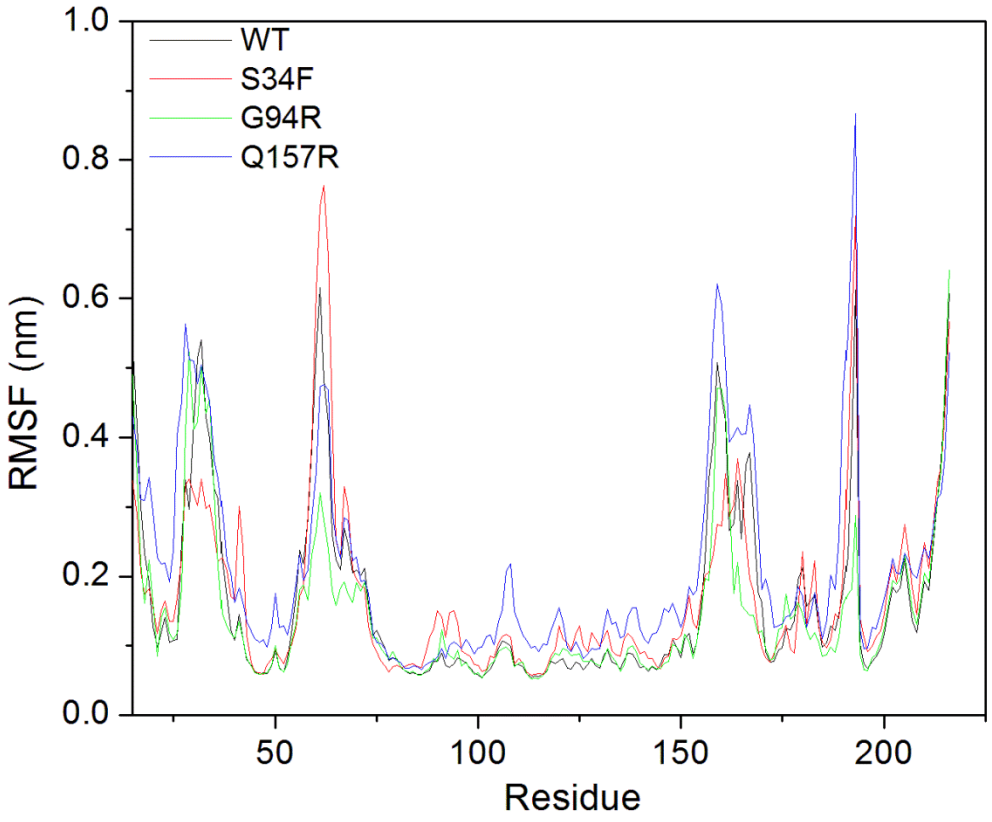


Figure 3: Figure represents the RMSF of atoms as a function of amino acids.

SASA analysis:

Evaluation of to what range an AA can interacts with the solvent and the protein core defines SASA. It is relative to the degree to which an AA is vulnerable to these environments [28]. Fluctuation in the SASA indicates the alteration in bared AA residues hence influencing the protein structure. SASA outcomes for simulations revealed a little bit of variation between the WT and the mutants shown in Figure 4. Among these mutants, Q157R showed a rise in the fluctuation from initial ns to 20ns which increases around 30ns followed by the equilibrium after 35ns. On other hand S34F has shown least variation, rise in the fluctuation initially followed by

the equilibrium around 20ns and then sudden fall in the fluctuation around 50ns further followed by the equilibrium. Whereas, G94R has shown no such fluctuations with respect to WT and is mainly at the equilibrium. Mutants S34F: 123.6, G94R: 123.6 showed lesser average total SASA compared to the WT: 124.16 and Q157R: 127.5 (Table 10).

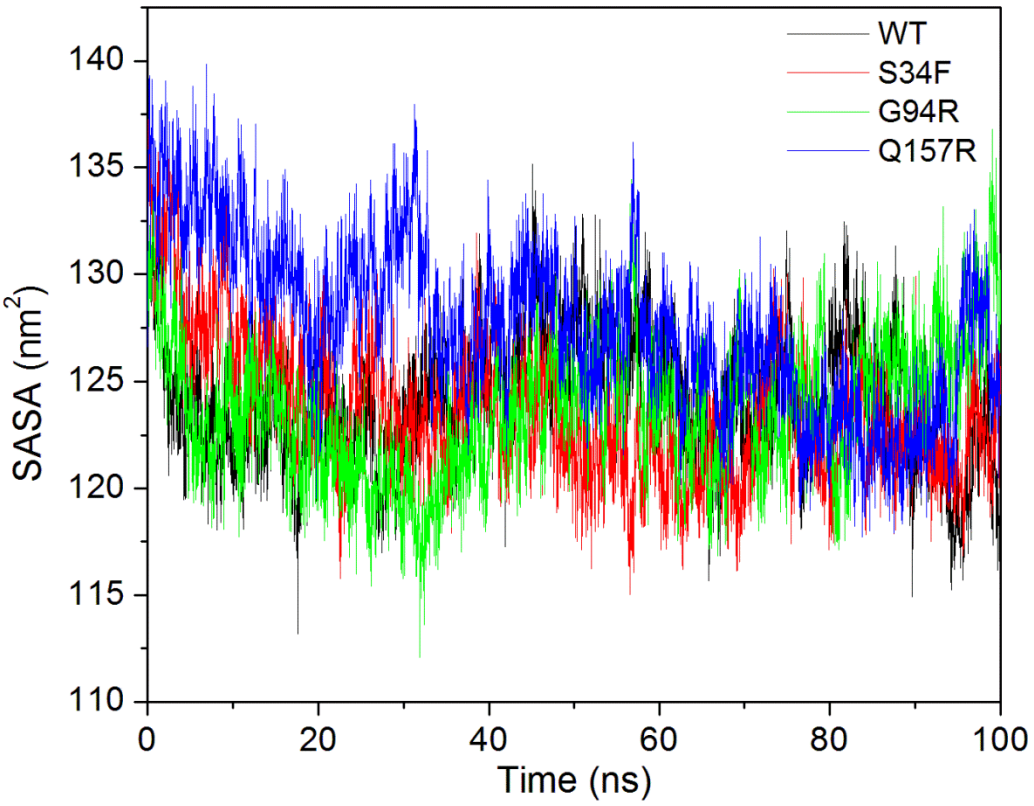


Figure 4: Figure represents the solvent-accessible surface area (SASA) of WT and MT versus time.

Radius of gyration analysis:

Radius of gyration is a measure to depict the poise configuration of a whole system specially in predicting the level of compactness in the structure of protein [29]. Rg plot for the protein and the atoms with time over the method of 100ns simulation is shown in Figure 5. We noted a prominent fluctuation in mutants in comparison with WT in this Rg plot. Among these MTs, G94R and S34F has not shown any large deviation with respect to the WT. A slight rise in the

fluctuation around 90ns is observed in the G94R simulation whereas; rise in the fluctuation is shown in the initial ns and around 90ns in S34F. Large range of fluctuation is shown bt Q157R as a sharp rise in the fluctuation from initially to around 20ns followed by the equilibrium at 50ns and slight rise around 90ns. Among these mutants, large deviation in Rg from the WT structure was witnessed during the simulation of Q157R (Table 10) (Figure 5). These outcome specify that as compared to other MTs, native protein could have experienced a major structural transition resulting from Q157R.

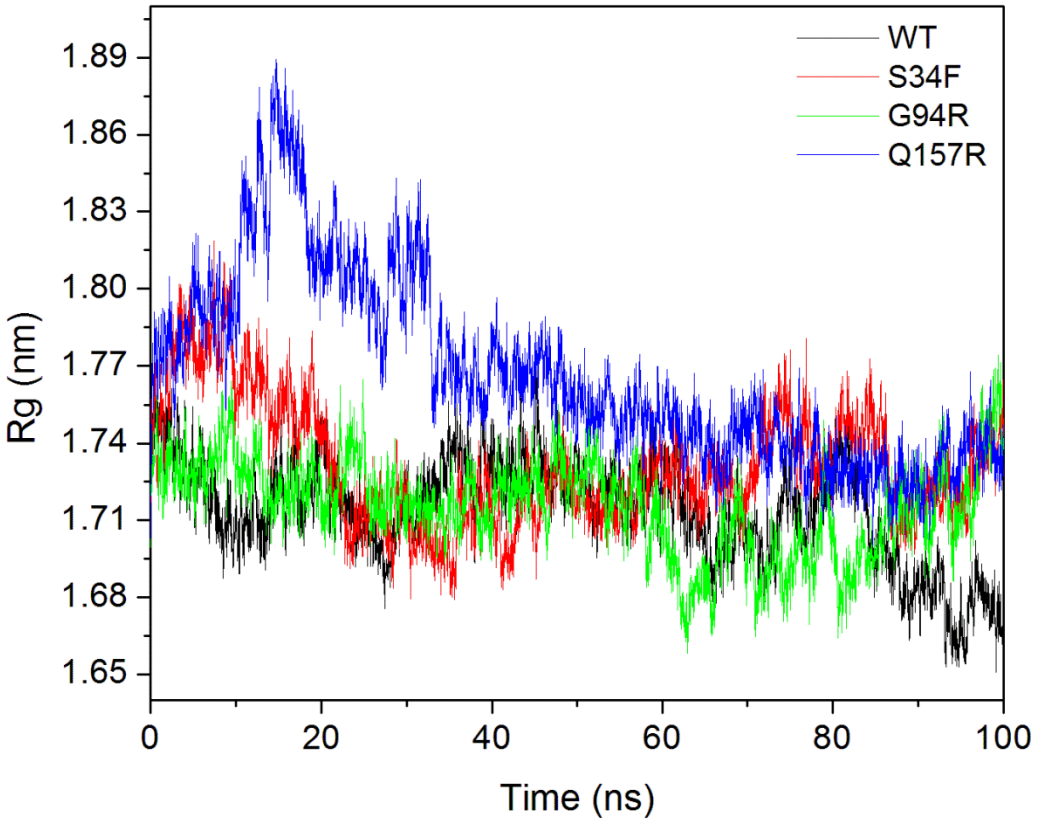


Figure 5: Radius of gyration of atoms during a 100ns MDS for WT and MT versus time.

Hydrogen-bonding analysis:

Another attorney that is important for preserving the stable conformation of a protein is hydrogen bonding. We have conducted the NH bond analysis of WT and MTs during simulation and plotted in the Figure 6 to comprehend the cause of flexibility between the mutants. Results

exhibited a slight difference in protein-solvent intermolecular hydrogen bond pattern within the WT and MTs. Among these mutants, a decrease around 20ns and increase in the number of hydrogen bonds was noted in G94R around 30ns. S34F has shown increase in the number of hydrogen bonds around 30ns. Whereas, Q157R has shown no variation either increase or decrease in the number of hydrogen bonds with respect to WT.

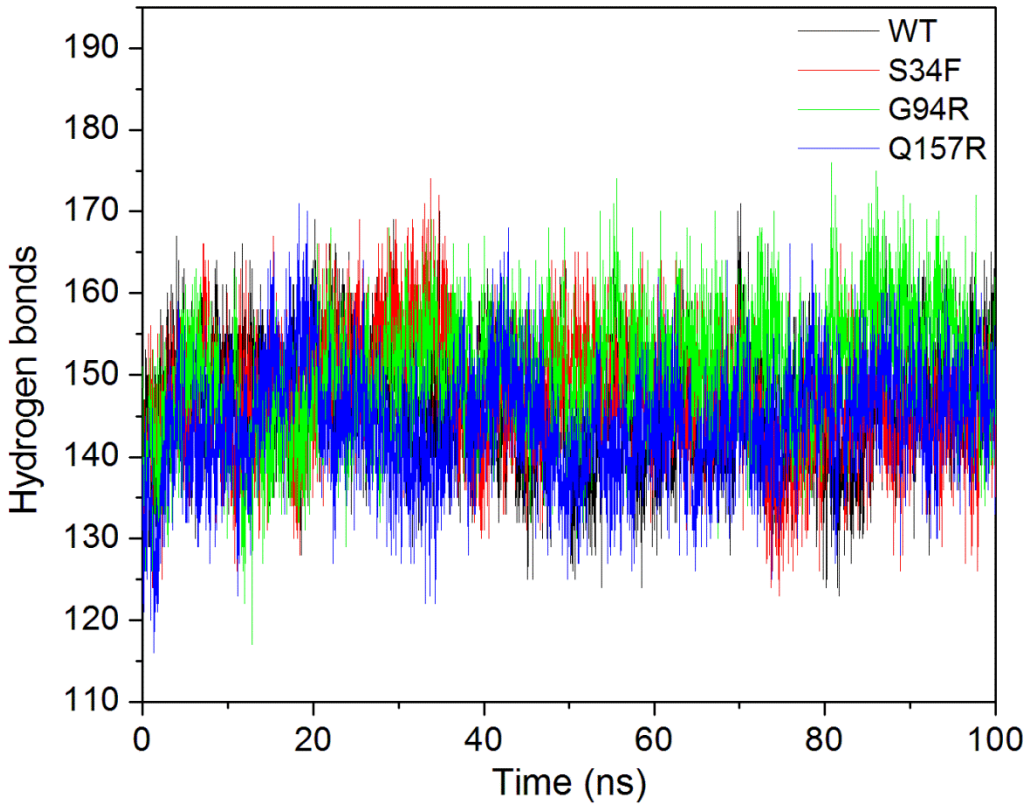


Figure 6: Figure represents the average number of protein-solvent intermolecular hydrogen bonds in WT and MT versus time.

CHAPTER 4

CONCLUSION

This project work will provide an understanding towards a genotype and phenotype relationship of deleterious lung cancer associated nSNPs in *U2AF1*. This study reports three mutations, Q157R, S34F, and G94R associated with lung cancer phenotype and the Molecular dynamic simulation showed the effect of these mutants on native protein. RMSD, RMSF, Rg, SASA, Hydrogen bonding uncovered their deleterious impact on the native protein. Therefore, this computational approach can provide an extensive notion to undermine the mechanisms of these SNPs in lung cancer. The conclusions stated in this project illustrate the function of deleterious mutations which can give beneficial information for the designing of these mutants based restorative approaches fighting lung cancer.