Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI:: 10.53555/jptcp.v30i7.7476

IN SILICO ANALYSIS OF NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS ASSOCIATED WITH FLT3 GENE

Salheen Bakhet[1*], Dr. Muhammad Shoaib[2], Taliah Tajammal[3], Dr. Iram Aziz[4]

[1*,2,3]Computer Science Department, University of Engineering & Technology, Lahore, Pakistan.
[4]School of Biological Sciences, University of the Punjab, Lahore, Punjab

[*]**Corresponding Author:** Salheen Bakhet
*Email:salheen@ieee.org

Abstract
SNPs play a vital and important role in the genetics and phenotype changes in humans. It is a genetic reason of complicated, critical and lethal diseases. In silico is referring to mass use of silicon. It is an expression which means performed on computer or via computer simulation. FMS like Tyrosine Kinase 3 or Fatal Liver Kinase 2 is a protein that in humans is encoded by the FLT3 gene. It has become significant to extract the facts how the functionality and molecular dynamic behavior of SNPs may affect genetic behavior of FLT3 gene. Emphasis is made in investigating pathogenic effects of nsSNPs in FLT3 gene using computational tools. A total of 638 missense SNPs were extracted from dbSNP of NCBI. These 638 missense SNPs were further processed through different layers and a final list of 9 most deleterious SNPs reported by almost every tool were identified.

Keywords: Flt3, Cancer, Leukemia, Sift, Condel, Snap2, Snp&Go, Provean, Panther, Phd_Snp, Polyphen2, Cadd, Cupsat, Mupro, I-Mutant

INTRODUCTION

Importance of SNPs is measured from the abnormal behavior which occurs in human body due to the changes concerned SNPs may cause in FLT3[1] gene and which may lead to deadly cancerous diseases such as Acute Myeloid Leukemia (AML) (Pelcovits et al., 2020) [2] and Chronic Myeloid Leukemia (CML) (Heim et al., 2020) [3]. This is the reason why scientists and researchers give high attention to this gene and to the relevant SNPs which may occur in the gene.

AML or CML is a complicated, dynamic human malignancy and fatal disease. It is characterized by multiple somatically-acquired driver mutations, poor prognosis and little survival. FLT3 is mentioned aberrantly in most of AML patients. The number of survived adult patients after 5 years who were diagnosed AML is almost 20%.

The association between the polymorphism and risk of AML had already evaluated. Recent meta-analysis by using computational tools suggests that polymorphism plays important role in the vulnerability of AML / CML. The study of investigation of patients involving AML and cytokine polymorphism is uncommon and irregular. In this research, computational analysis of non-synonymous SNPs (nsSNPs) in FLT3 gene was performed so that all possible deleterious mutations might be identified. A proposition for modeled structure of mutant protein might also be proposed. A molecular dynamic simulation using GROMACS (Spoel et al., 2005) [4] was also designed.

It is believed that the results of the research will provide an insight and further understanding about human diseases in FLT3 gene. It is also believed that the same may also prove a guide for future experimental work.

1. CANCER

Cancer refers to any of the many diseases that are characterized by the formation of abnormal cells that divide uncontrollably and have the ability to invade and destroy normal body tissues. Cancer is the second leading cause of death in the world.

There are almost more than 100 types of cancers reported till the date. One of the common reasons for cancer to form is a deleterious SNP. Some of the cancers can be cured at early stages but later in advance stages cure is not possible. For every type of cancer, scientists categorize the stages based upon the type of cell abnormality and their growth: Stage 0 states that there is no cancer, only abnormal cells have the potential to become cancerous. Stage I means the cancer is small and is only in one place. Stages II and III mean that the cancer is large and has grown into nearby tissues or lymph nodes. Stage IV means that the cancer had spread to other parts of the body. See (Figure 1)[5]

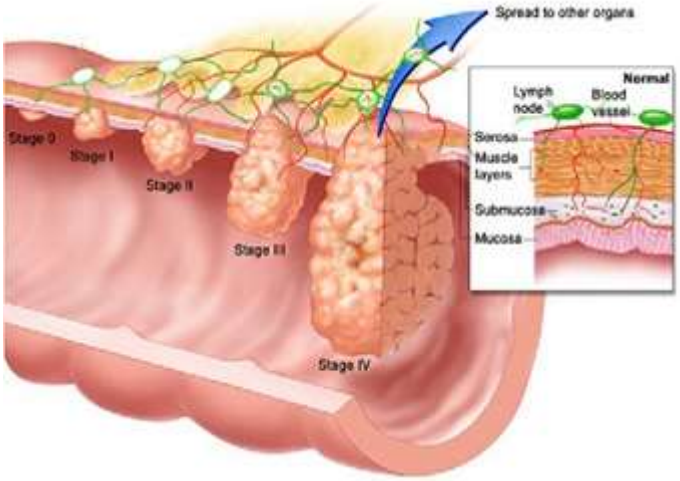

Figure 1 - Depiction of all five stages of cancer [5]

Breast, prostate, lungs, colon & rectal and melanoma cancer are the top five types of cancer reported every year worldwide. Some cancers, such as leukemia, do not form tissues. The blood cells made from bone marrow are of several types; red blood cells, neutrophil, lymphocytes, monocytes, basophil and platelets. The primary function of red blood cells (erythrocytes) is to carry oxygen from lungs to all parts of the body. White blood cells which are also called as leukocytes involved in fighting off infections and empower body's immune system. Platelets (thrombocytes), the smallest blood cells, on the other hand involved in the blood clotting to heal the bleeding. (Figure 2) [6] illustrates the shapes of all the blood cell types.

Three major types of blood cancer are:

(i) Leukemia
(ii) Lymphoma
(iii) Myeloma

Leukemia is related to uncontrolled growth of blood-forming tissues which might occur in bone marrow or in lymphatic system. Leukemia is equally common in children and adults. It mostly involved in the white blood cells. The major disease is known as acute myeloid leukemia. More than 60,000 people are diagnosed with leukemia each year.

Lymphoma is a type of blood cancer starts in infection-fighting cells present in the immune system of a living organism.

Myeloma was another type of cancer forms in plasma cells and it made the cells unable to form antibodies so the cells cannot recognize and attach to germs or bacteria. Leukemia is the most common and most studied type of blood cancer.

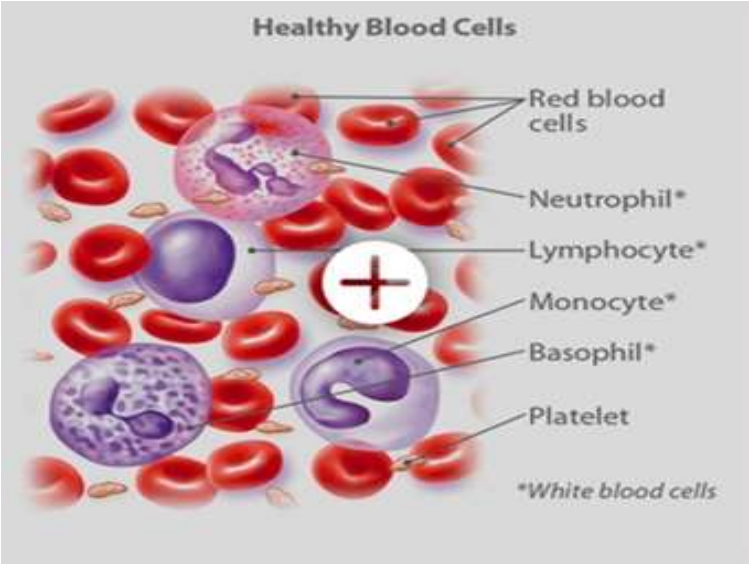

Figure 2 - Depiction of healthy blood cells [6]

There are four main types of leukemia. These are classified as acute or chronic, and myeloid or lymphocytic:

(i).Acute myeloid (or myelogenous) leukemia (AML)
(ii).Chronic myeloid (or myelogenous) leukemia (CML)
(iii).Acute lymphocytic (or lymphoblastic) leukemia (ALL)
(iv).Chronic lymphocytic leukemia (CLL)

1.1. BLOOD CANCER TREATMENT AND THERAPY OPTIONS

Treatment for blood cancer depends on the type of cancer, age, how fast the cancer is developing, where the cancer has expanded and other factors. Other therapies for leukemia include:

(i).STEM CELL TRANSPLANTATION
(ii).CHEMOTHERAPY
(iii).RADIATION THERAPY

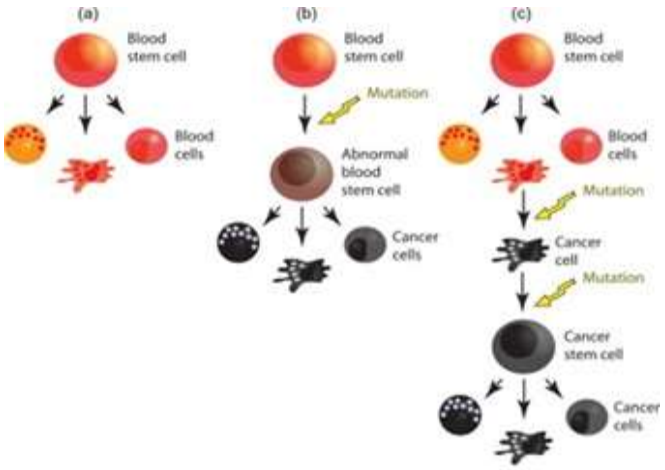

Figure 3 - Normal & cancerous blood stem cells [7]

1.2. RECEPTOR TYPE TYROSINE-PROTEIN KINASE FLT3

FLT3 is abbreviation of FMS-like Tyrosine Kinase 3. Antigen 135 (CD135) cluster of differentiation is also known as FMS-like Tyrosine Kinase 3. The FLT3 gene is located on chromosome 13 (13q12). Both the "Receptor type tyrosine-protein kinase FLT3" and the "Fatal-liver Kinase-2 (FLK2)" are proteins in human encoded by FLT3 gene.
The following two different types of FLT3 gene mutation are diagnosed in AML cases.

(i).Internal tandem duplication (ITD) mutations, which occurs within the juxtamembrane region of the gene;
(ii).Point mutations which happen at codon835(D835) within the kinase domain.

Both types of mutations partially activate the actions of tyrosine kinase. The incorrect diagnostic and strong leukocytosis are strongly related and associated with mutations in FLT3 gene. Such mutation has high frequency of reoccurring again and again in any of the patients recovered from it. The survival rate of its patients is low.
Cytokines (interleukins [ILs], growth factors, interferons, etc.) plays vital role in regulating the inflammatory response. These also play important role in chronic inflammation and are involved in the growth and development of cancer. Chronic inflammation relates to the release of various mediators (proinflammatory and oncogenic ones). These may include reactive nitrogen oxygen species, inflammatory cytokines (IL-1β, IL-2, IL-6, and tumor necrosis factor alpha [TNF-α]), growth factors and chemokines.

2. MATERIALS AND METHODOLOGY

The analysis of SNPs and its deleteriousness are primarily divided into three major categories:

(i).Deleterious / disease associated behavior
(ii).Protein stability
(iii).Pocket identification

A total number of 13 computational tools were used under each of the categories to produce results. A number of 9 different tools were used to identify SNPs with deleterious disease associated behavior, a number of 3 different tools were used for protein stability and a tool was used for pocket identification. The following tools were used to identify SNPs with deleterious / disease associated behavior:

(i). SIFT
(ii). CONDEL
(iii). SNAP 2
(iv). SNP & GO
(v). PROVEAN
(vi). PANTHER
(vii). PHD_SNP
(viii). POLYPHEN 2
(ix). CADD

The following tools were used for protein stability:
(i).CUPSAT
(ii).MuPro
(iii).I Mutant

The following tool was used for pocket identification:
(i). Castp

2.1. DATASET DOWNLOADED

Sequence of Homo sapiens FLT3 gene (accession number NC_000013.11) was retrieved from online OMIM database with FLT3_Human entry code P36888. Details of Single Nucleotide Polymorphisms (SNPs) obtained through NCBI dbSNP repository http://www.ncbi.nlm.nih.gov/snp. A total of 638 missense SNPs were extracted from a total of 24784 SNPs. The data was further sort out where rs_IDs, single nucleotide variance (SNV) and SNP position on genome assembly (GRCh37 and GHCh38) versions were extracted. Several tools were used to find the deleterious / disease associated SNPs from the missense SNPs. Protein stability and the effect of these dreaded and deleterious SNPs on protein stability were also examined.

2.2. SNP DISEASE ASSOCIATION PREDICTION

The outcome of every SNP disease association prediction tool will be explained.

2.2.1. SIFT

SIFT (Vaser & Adusumalli et al., 2015) [7] stands for **S**orting **I**ntolerant **F**rom **T**olerant. This tool is used to predict that amino acid replacement affects protein performance so users can prioritize substitutes for further learning. It uses homologous sequences to predict amino acid replacement which will affect protein function and therefore, may alter the phenotype. Considering that disease-causing amino acid substitutes are detrimental to protein performance, we used SIFT in the missense substitute database related to disease involvement. It was used in a variety of human data and was able to distinguish mutations associated with disease from neutral polymorphisms. The tool reported, with an accuracy of 89.81%, all the SNPs as deleterious that had score < 0.5. Here it is notable that the highest accuracy was provided by the results of SIFT tool. The results and values for SIFT were produced using CADD tool.

2.2.2. CONDEL

CONDEL (Gonza´lez-Pe´rez et al., 2011) [8] is abbreviation of **CON**sensus **DEL**eteriousness. It is a combination of various tools. It represents score of non-synonymous single nucleotide variants (SNVs) and is a method to assess the outcome of non-synonymous SNVs. It was originally developed to integrate the outputs of five different tools such as SIFT, MAPP, PolyPhen 2, Mutation Assessor and LogR Pfam E-value. The CONDEL score weighs different methods using complementary cumulative distributions of approximately 20 deleterious and neutral missense SNPs. The probability that an unpredictable change in status is not a positive factor and that the probability that a predictable neutral change is not a false negative was used as a weight. The tool reported, with an accuracy of 88.40%, all the SNPs as deleterious that had CONDEL score > 0.6.
It gave second highest accuracy as compared to the other tools used for the purpose.

2.2.3. SNAP 2

SNAP2 (Hecht et al., 2014) [9] is a neural network based classifier which was trained over 100,000 variants from OMIM repository and a set of pseudo-neutral variants. Interpretation of the results of the broader genome organization is confused with the linking of disease between adjacent pathways. The SNAP2 is a dynamic bioinformatics query tool for single-nucleotide polymorphisms (SNPs) to identify and interpret SNPs involved in linking disequilibrium (proxies) based on HapMap. The tool receives a substitution as amino acid input and a score which reflects the likelihood of a possible mutation. The SNAP server facilitates the interpretation and comparison of genome-wide relationship research results, as well as the development of good map testing (by defining genomic regions with variability relative to their proxies).
SNAP2 is known to be one of the frequently used and user-friendly tools with highest accuracy of 82%. The tool reported all the SNPs as deleterious that had score > 60%.

2.2.4. SNP&GO

SNP&GO (Majumdar et al., 2017) [10] is a web server for predicting single-point protein mutations related to human diseases. It is a tool used to detect harmfulness of a SNP by exploiting the corresponding protein functional annotation. Widely investigated SNPs are missense mutations that lead to the insertion of residue into proteins. SNPs&GO is an accurate method proposed here. It is based on support vectore machine algorithm. It helps predicting disease-related mutations from protein sequences, scoring points with 81% accuracy. The genetic basis for human diversity is mainly due to Single Nucleotide Polymorphisms (SNPs). SNPs & GO collects unique framework information found in protein sequence, protein sequence profile, and protein activity. Its accuracy for the value of P was higher compared to both PhD_SNP and PANTHER tools (to be described later).

The tool reported all the SNPs as deleterious that had values of $P > 0.5$.

2.2.5. PROVEAN

PROVEAN (Choi et al., 2015) [11] stands for **PRO**tein **V**ariation **E**ffect **AN**alyzer. This is another tool used to measure the effects of amino acid substitutions on the biological function of protein activity. Its performance is commendable and comparable with popular tools such as SIFT and PolyPhen 2. Its scores are calculated based on homologs collected in the database. PROVEAN is able to provide an estimate of any type of protein variation as follows:

- Single or multiple amino acid substitutions
- Single or multiple amino acid insertions
- Single or multiple amino acid deletions

It proved quite useful for classifying sequence variants for the identification of non-synonymous variants. The tool is quite popular and valued when comparing to SIFT and PolyPhen 2. The tool reported, with an accuracy of 79.50%, all the SNPs as deleterious that had score < -2.5.

2.2.6. PANTHER

PANTHER (Thomas et al., 2003) [12] is abbreviation of **P**rotein **AN**alysis **TH**rough **E**volutionary **R**elationships. It is a method to relate protein sequence relationships to function relationships in a healthy and valid way. Proteins are classified as follows:

- Family and subfamily
- Molecular function
- Biological process
- Pathway

It provides a classification system and a large database of genetic and protein families as well as small functional families or sub-families that are used to classify and identify the function of genetic products. Its segregation is the result of human curation and bioinformatics algorithms. Its APIs provides its services which are designed to operate independently and are part of a larger workflow.

The tool reported, with an accuracy of 79%, all the SNPs as deleterious that had $P > 0.5$. It provided 1% more accuracy as compared to PhD_SNP.

2.2.7. PHD_SNP

PhD_SNP (Tian et al., 2007) [13] is abbreviation of **P**redictor of **H**uman **D**eleterious **S**ingle **N**ucleotide **P**olymorphisms. It uses a predictor according to a single SVM trained and tested for sequence of proteins and profile details. It is evident from the name that the tool predicts deleterious SNPs in human body.

The PhD_SNP SVM input was primarily built using the following steps:

- In the given variation the filing form for the rest of the fossil record was encoded in a 20 vector with 1 in the position relating to the fossils, 1 in the area relating to the fossils and 0 in the remaining 18 positions.
- The 20-element vector encoding of its sequence form created a report of the appearance of fossils in 19 fossil windows around the converted fossils.
- With a given protein, its sequence was constructed according to the process described above. From this we examined both the wild type (Fi (WT)) and the modified fossils (Fi (MUT)) in the position i. The NAL number was the sequence of the given position and position and the Conservation Index (CI).

The approach of the tool relies upon Single Vector Machine (SVM) based classifier. The tool reported all the SNPs as deleterious that had values of P > 0.5. It provided accuracy of 78%.

2.2.8. POLYPHEN 2

PolyPhen 2 (Adzhubei et al., 2015) [14] is abbreviation of **POLY**morphism **PHEN**otyping version 2. This tool is used to predict the possible emergence of amino acid synthesis in the structure and functions of human proteins. The substitution was made following straightforward physical and proportional considerations. Some of the highlights of the new features of the tool are as follows:

- High quality sequence alignment pipe.
- Possible distinctions are based on machine learning method.
- It is designed for high-level sequencing data analysis for the next generation.

It was shown in several studies that the impact of amino acid allelic variability on protein / activity formation can be accurately predicted by multiple sequential correlation analysis and 3D structures. These predictions were consistent with the effect of natural selection which was seen as skipping unusual alleles. Thus, cell-level prediction revealed SNPs that affected actual phenotypes.
The tool reported, with an accuracy of 76%, all the SNPs as deleterious that had score > 0.5.

2.2.9. CADD

CADD v 1.6 (Rentzsch et al., 2018) [15] is used to produce all the results. CADD is the only popular single combined tool used to measure heterogeneous annotations. It uses data and statistics from the Ensembl Variant Effect Predictor (VEP), the ENCODE project and the UCSC genome browser tracks. Besides all this, it also uses data types like GERP, PhastCons, PhyloP, DNase hypersensitivity, transcription factor binding and protein level scores like SIFT, PolyPhen, etc. A C_score with a higher or equal score of 10 indicates an expectation of 10% scary motives you could do to human genes, 20 or more equal points showed 1% worse and so on. CADD is commonly used to be reliable tool because it was created by accumulating the methodologies and the best practices of many other tools. One of the most important benefits of CADD tool are that it provides a platform to produce results of other tools such as PolyPhen 2, SIFT, etc.
The tool reported, with an accuracy of 74.30%, all the SNPs as deleterious that had C_score > 20.

2.3. SNP IMPACT ON PROTEIN STABILITY

With the systematic approach, we had first found deleterious and disease associated behavior. The next task was to find out SNP impact on protein stability. Three popular tools of high-precision was selected. The highest accuracy had been noticed for CUPSAT tool. An explanation of the results will be provided with three tools that will be explained next.

2.3.1. CUPSAT

CUPSAT (Parthiban et al., 2006) [16] is a web-based tool for analyzing and predicting protein changes in point mutation (single amino acid modification). It has the following key features:

- CUPSAT uses natural protein strength to predict protein state (stabilized / destabilized).

- The power of the amino acid-atom is used. With this, 40 types of amino acids from Melo-Feytmans are used to enhance radial pair diffusion activity.
- Torsion angle strength is obtained from the distribution of large torsion φ and ψ angles.
- The function of Gaussian apodization has been used to increase torsion angle perturbation in protein mutants.
- Mutant strength assumptions from PDB and custom protein structures are possible.

CUPSAT had shown highest accuracy of 87% amongst all the other tools used for measuring protein stability. More the value got smaller than -0.5, higher was deleteriousness of the SNP for protein stability.

2.3.2. MUPRO

MuPro (Cheng et al., 2006) [17] tool is comprised of multiple machine learning techniques to assess and find out as how mutation in single amino acid affected protein stability. Two main methods used for machine learning algorithms are based on Support Vector Machines (SVM) and Neural Networks. The tool was trained with large mutation datasets and validated with 82% or more accuracy which was better compared to other tools in the literature. Any smaller value(s) less than -0.5 was considered to be the most deleterious for protein stability.

2.3.3. I-MUTANT

I-Mutant is a suite used to predict the effects of single-point protein conversions / mutations. The tool is based on Support Vector Machines (SVMs) algorithms which are deployed and hosted at a distinctive web server. It provides the ability to automatically predict changes in protein density at a single site conversion starting with a single protein sequence or protein structure where available. The tool had given lowest accuracy of 80% for the selected SNPs for protein stability. Any smaller value(s) less than -0.5 was considered to be the most deleterious for protein stability.

2.4. POCKET IDENTIFICATION

Pocket identification is the technique to find the binding site of protein for ligands and adjacent proteins. Computational methods are used to find the pocket on probability based algorithm trained by Artificial Neural Network (ANN) protocols. CastP, a web-based tool was used for the pocket identification and residues present in first pocket/active site/binding was selected for docking analysis.

2.5. PROTEIN MODELING AND PROTEIN LIGAND DOCKING

Till the date, there are seven PDB structures of FLT3 encoded protein 'Receptor-type tyrosine-protein kinase' present in Protein Data Bank (https://www.rcsb.org). Among them, PDB ID: 4rt7 is downloaded with maximum length, containing the kinases domains in it. Ligands are removed from the structure in order to add mutations in the original protein structure (Vallejos-Vidal et al., 2020) [18]. By using FoldX (Bub et al., 2018) [19], we incorporate mutations and save the structures separately. Original and Mutant structures are then superimposed in order to see the structural variations caused by the mutations. The already attached ligand of original 4rt7 protein was again docked with the mutants and found the interactions bond difference caused by the mutations in protein structure. The variants on the residues bind to the ligand were further investigated and selected mutants were evaluated with their Root Mean Square Deviation (RMSD) using Molecular Dynamics Simulation. GROMACS was used to find out the RMSD values over the period of 20ns.

3. RESULTS

SNPs work as a marker in order to help scientists to find out genes which are associated with the disease.

No.	SNP IDs	No.	SNP IDs
1	770630954	39	1280596886
2	568745490	40	748106520
3	1057519762	41	1403916427
4	1216374794	42	746946336
5	779366241	43	768037348
6	1338788116	44	1224002756
7	1221467960	45	1276477642
8	1465068581	46	974980406
9	772146579	47	1264826559
10	1057519767	48	142796469
11	1057519768	49	1337282620
12	770301265	50	868802040
13	1361689773	51	368432815
14	1428096626	52	774360726
15	1327136575	53	903856095
16	201504848	54	376588714
17	773272572	55	1403972292
18	772061268	56	1221240514
19	1452571288	57	769394501
20	201923726	58	201208287
21	764151292	59	773758373
22	1322506485	60	1474119276
23	200909894	61	750415487
24	149791635	62	375798213
25	1057520025	63	774511256
26	754396362	64	1423397744
27	866751216	65	776133815
28	1268548071	66	762198688
29	967108282	67	746926889
30	1423011599	68	780941129
31	370459694	69	1442266284
32	769640419	70	1057519769
33	1244508922	71	751544883
34	1227747438	72	1207508622
35	776386970	73	779688631
36	776952835	74	1275470574
37	758095462	75	1312380044
38	1367156928		

Table 4.1 – List of missense SNPs with high probability.

When a SNP occurs in a gene or in a regulatory region near a gene, it plays more direct role in diseases by affecting the gene's function. However, it is also true that many SNPs do not have any effect over human health. In order to figure out deleterious effects of SNPs in FLT3 gene and its regulatory regions, a fruitful exercise was conducted to diagnose deleterious and disease associated results through the aforementioned computational tools. Moreover, emphasis was also made over protein functionality related amino acid substitutions for controlling disease causing effects of SNPs. **Table 4.1** provided a list of 75 deleterious SNPs. The results were primarily categorized based on the outcome of deleterious / disease associated and protein stability computational tools. Out of a list of 24784, a total of 638 missense SNPs were diagnosed. These missense SNPs were

further input one by one to each of the 12 disease associated and protein stability computational tools. Different tools interpreted its scored results differently. E.g. tools such as SIFT, CADD and PROVEAN interpreted its scored results for SNPs as deleterious, tools such as PHD_SNP, PANTHER and SNP&GO interpreted its scored results as disease and similarly other tools interpreted its results as favorable / unfavorable or stabilizing / destabilizing. All computational tools reported different number of deleterious SNPs.

SIFT, CONDEL, SNAP2, SNP_GO, PROVEAN, PANTHER, PHD_SNP, POLYPHEN, CADD, CUPSAT, MUPRO and I_MUTANT reported different number of disease associated and protein destabilizing SNPs. A number of 23, 25, 26, 30, 40, 43, 44, 50, 59, 65, 73 and 75 deleterious SNPs were reported by PANTHER, CUPSAT, CONDEL, SNP_GO, PHD_SNP, I_MUTANT, MUPRO, SNAP2, PROVEAN, CADD, POLYPHEN and SIFT respectively. From the set of deleterious and disease associated tools, PANTHER reported lowest number of SNPs whereas SIFT reported highest number of SNPs. From the set of protein stability tools, MUPRO reported highest number of deleterious SNPs whereas CUPSAT reported lowest number of deleterious SNPs.

Only SIFT reported 75 missense SNPs out of which 23 SNPs were commonly reported to be possible deleterious by all the other computational tools. The most and the least numbers of deleterious SNPs were reported by SIFT and PANTHER respectively. Each set of disease association and protein stability tools reported an average of 37 deleterious SNPs. It is analyzed that SIFT reported superset list of SNPs which was one way or the other reported by all the other tools. Three kinds of lassifications were discovered in the initial analysis. PANTHER reported lowest number of 23 possible deleterious SNPs. An additional upto 52 probably deleterious SNPs were reported by other computational tools. Thus, it was concluded if SNP occurred at any of the 75 positions – there was a probability of AML. If 37 (average) SNPs occurred, there was a higher probability of AML. Similarly if any of the most common 23 SNPs reported by all the tools occurred, such SNPs were the most deleterious and there was highest probability of AML.

Upon detailed analysis, it further became evident that there were 9 SNPs, listed in **Table 4.2,** equal or greater than the cutoff conditions of most of the computational tools. These SNPs prominently emerged as the most qualified, highly and possibly deleterious compared to all other SNPs. Thus, it was concluded that AML was guaranteed if any of these 9 SNPs occurred.

Tool	**Condition**	**SNP IDs**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
		1057519768	**770301265**	**201923726**	**754396362**	**967108282**	**1423011599**	**370459694**	**758095462**	**1224002756**
SIFT	Score < 0.5	0	0	0	0	0.01	0	0	0	0
CONDEL	Score > 0.6	0.562926961	0.68689459	0.60075314	0.52955328	0.65330691	0.618930876	0.5607622	0.53986914	0.729812893
SNAP 2	Score > 60%	48	72	81	15	84	65	78	11	68
SNP & GO	P > 0.5	0.763	0.847	0.793	0.645	0.861	0.838	0.755	0.772	0.848
PROVEAN	Score < -2.5	-7.469	-8.404	-11.452	-7.077	-8.29	-8.167	-4.536	-7.829	-9.771
PANTHER	P > 0.5	0.583	1	0.581	0.725	0.684	0.996	0.986	0.753	0.684
PHD_SNP	P > 0.5	0.834	0.936	0.887	0.773	0.885	0.898	0.853	0.913	0.947
POLYPHEN2	Score < 0.5	0.91	1	0.999	0.982	0.911	1	1	0.989	1
CADD	Score < 20	27.2	25.3	27.2	28	25.8	23	22.4	29.9	21.8
CUPSAT	ΔΔG<0 High Impact ΔΔG<-0.5	1.32	-6.82	3.99	2.41	2.58	-2.81	5.88	-0.72	3.39
MuPro	ΔΔG<0 High Impact ΔΔG<-0.5	-1.182	-0.526	-0.906	-0.62	-1.45	-1.007	-1.47	-0.747	-0.828
I Mutant	ΔΔG<0 High Impact ΔΔG<-0.5	-1.12	-0.87	-0.78	-0.58	-1.39	-0.97	-1.29	-0.652	-0.797

Table 4.2 – List of highly deleterious 9 SNPs reported by all the tools.

Some of these tools reported bordering results while significantly lower results were reported by other tools. Therefore, described 9 SNPs were required to be dealt very carefully and sensitively in order to avoid, treat or monitor AML.

3.1. POCKET IDENTIFICATION, PROTEIN MODELING AND MOLECULAR DYNAMIC SIMULATION

We further investigated active site residues among these nine residues using CASTP and found that four mutants (C694R, C828Y, Y842C, and Y842H) at amino acid level were involved in the protein binding site. See (Figure 4)

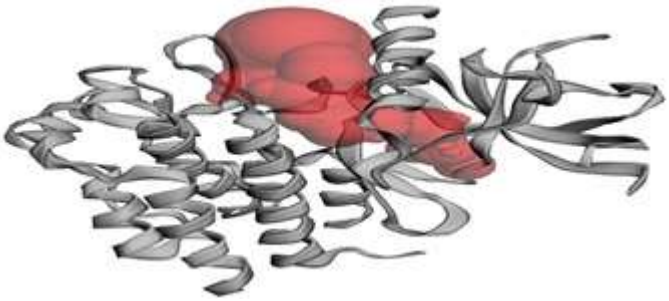

Figure 4 - depicted active site of FLT3 protein in red balls whereas rest of the structure is in cartoon shape.

We then downloaded the most verified X-ray diffracted structure (4rt7) from Protein DataBank (Figure 5).

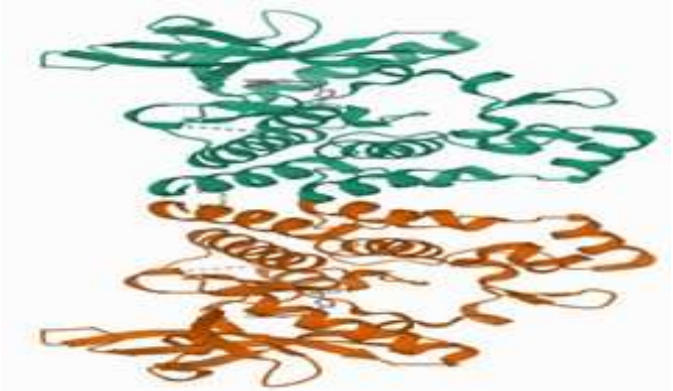

Figure 5 - showed the original structure of ID: 4rt7 downloaded from Protein Databank**.**

The resolution of the model was 3.10Å with Ramachandran outliers 0.8%. We then refined structure through FoldX and built mutant models (Figure 6)

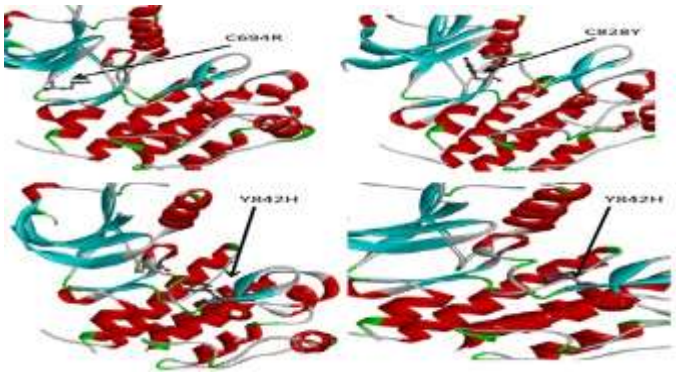

Figure 6 - depicted the mutations at the selected residue sites.

As the selected variants were structurally buried so they did not show any major change on the surface of the protein.

The original structure was opened using LigPlot+ (Wallace et al., 1995) [20] in order to see the protein ligand interaction and we found that two residues Cys694 and Cys828 were directly involved in the binding with ligand (P301001) as shown in (Figure 7).

Figure 7 - Protein-Ligand interaction of original structure depicted the interactions of our candidate residues with ligand.

Molecular Dynamics Simulation of these mutant PDB structures (built from FoldX) was done in order to see the RMSD from the native structure.

Properties	**WT pVHL**	**C694R pVHL**	**C828Y pVHL**
Backbone RMSD (nm)	0.1282 (0.02)	0.1527 (0.02)	0.1444 (0.02)
Cα RMSD (nm)	0.1332 (0.02)	0.1584 (0.02)	0.1506 (0.02)
Rg-protein (nm)	1.2912 (0.06)	1.2861 (0.08)	1.2800 (0.06)

Table 4.3 – Time averaged structural properties calculated for Native, C694R and C828Y pVH. Standard deviation values were given in parentheses.

MD simulation revealed us that there was a variation in both variants as compared to the wild type structure. C694R showed a greater fluctuation upto the 8000 ps (8 ns) whereas C828Y was following the trend of native structure. However later all the structures attained more or less same pattern in a straight way. The radius of gyration (Rg) of native structure showed a fluctuation between 1.2684 to 1.3169 nm. C694R and C828Y showed the Rg value 1.2601 and 1.3228 nm and 1.2579-1.3067 nm respectively. The overall pattern showed that these variants were providing the fluctuations in the atomistic molecular movements over the period of time. However these fluctuations were not so much significant so we could not rely on the results only instead we needed to see the behavior of these mutants through gene annotation in Next Generation Sequencing Technique.

The study on the pathogenic effects of nsSNPs in the FLT3 gene using computational tools shares a foundational reliance on in silico methods similar to those employed in network optimization and security frameworks. T. A. Khan et al.'s [21] work on topology-aware load balancing in datacenter networks exploration underscore the importance of computational methodologies in addressing complex problems, whether in genetics or network management. In the context of SNP analysis,

previous studies have shown the effectiveness of computational approaches in handling large-scale genetic data [22]. Similarly, advanced machine learning techniques have been applied in various domains, such as optimizing lending risk analysis and transforming agriculture through plant health monitoring [23].

4. CONCLUSION

FLT3 gene is one of the most important genes of human body. Highly deleterious nsSNPs within the gene may lead to dolorous and fatal cancerous diseases. There are other studies available describing spread of AML or CML within human body. But there is lack of analysis using as many computational tools as used in this study. This study presented a comprehensive analysis over considerable number of SNPs using various credible computation tools. It not only predicted disease associated behavior but also analyzed protein stability measures. The results analyzed and classified SNPs into non-deleterious or probably deleterious or highly possibly deleterious. A final conclusive 9 deleterious SNPs were identified, if occurred, were expedite the process of AML or CML within human body.

5. REFERENCES

1. Wu, M., C. Li, and X. Zhu, FLT3 inhibitors in acute myeloid leukemia. J Hematol Oncol, 2018. 11(1): p. 133.
2. Pelcovits, A. and R. Niroula, Acute Myeloid Leukemia: A Review. R I Med J (2013), 2020. 103(3): p. 38-40.
3. Heim, D., M. Ebnother, and G. Favre, [Chronic myeloid leukemia - update 2020]. Ther Umsch, 2019. 76(9): p. 503-509.
4. Van Der Spoel, D., et al., GROMACS: fast, flexible, and free. J Comput Chem, 2005. 26(16): p. 1701-18.
5. Patrick, C., Colon Cancer. 2020.
6. Markman, M., Blood cancers. 2021.
7. Vaser, R., et al., SIFT missense predictions for genomes. Nat Protoc, 2016. 11(1): p. 1-9.
8. Gonzalez-Perez, A. and N. Lopez-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet, 2011. 88(4): p. 440-9.
9. Hecht, M., Y. Bromberg, and B. Rost, Better prediction of functional effects for sequence variants. BMC Genomics, 2015. 16 Suppl 8: p. S1.
10. Majumdar, I., I. Nagpal, and J. Paul, Homology modeling and in silico prediction of Ulcerative colitis associated polymorphisms of NOD1. Mol Cell Probes, 2017. 35: p. 8-19.
11. Choi, Y. and A.P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics, 2015. 31(16): p. 2745-7.
12. Thomas, P.D., et al., PANTHER: a library of protein families and subfamilies indexed by function. Genome Res, 2003. 13(9): p. 2129-41.
13. Tian, J., et al., Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics, 2007. 8: p. 450.
14. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, 2013. Chapter 7: p. Unit7 20.
15. Rentzsch, P., et al., CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res, 2019. 47(D1): p. D886-D894.
16. Parthiban, V., M.M. Gromiha, and D. Schomburg, CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res, 2006. 34(Web Server issue): p. W239-42.
17. Cheng, J., A. Randall, and P. Baldi, Prediction of protein stability changes for single-site mutations using support vector machines. Proteins, 2006. 62(4): p. 1125-32.
18. Vallejos-Vidal, E., et al., Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (Salmo salar). Front Genet, 2019. 10: p. 1406.

19. Buss, O., J. Rudat, and K. Ochsenreither, FoldX as Protein Engineering Tool: Better Than Random Based Approaches? Comput Struct Biotechnol J, 2018. 16: p. 25-33.
20. Wallace, A.C., R.A. Laskowski, and J.M. Thornton, LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng, 1995. 8(2): p. 127-34.
21. T. A. Khan, M. S. Khan, S. Abbas, J. I. Janjua, S. S. Muhammad and M. Asif, "Topology-Aware Load Balancing in Datacenter Networks," 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), Bandung, Indonesia, 2021, pp. 220-225, doi: 10.1109/APWiMob51111.2021.9435218.
22. Nuthalapati, Aravind. (2022). Optimizing Lending Risk Analysis & Management with Machine Learning, Big Data, and Cloud Computing. *Remittances Review*, 7(2), 172-184. https://doi.org/10.33282/rr.vx9il.25
23. Nuthalapati, Suri Babu. (2022). Transforming Agriculture with Deep Learning Approaches to Plant Health Monitoring. Remittances Review, 7(1), 227-238. https://doi.org/10.33282/rr.vx9il.230